THREE ESSAYS IN THE ECONOMICS OF EDUCATION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Ben Safety Ost

August 2011

THREE ESSAYS IN THE ECONOMICS OF EDUCATION

Ben Safety Ost, Ph.D.

Cornell University 2011

This dissertation is a compilation of three essays.

The first essay uses longitudinal administrative data on teachers to investigate the relative productivity benefits of acquiring general versus task-specific human capital. Within a school, elementary teachers frequently change grade assignments and I exploit the resulting variation in grade-specific tenure to separately identify the effect of general teaching experience and specific experience. Using a value-added model that controls for teacher fixed effects, I find that both general experience and grade-specific experience improve teacher performance. In addition to providing evidence that the productivity returns to human capital can be sensitive to seemingly small changes in task requirements, this study furthers our understanding of how teachers improve with experience.

The second essay uses longitudinal administrative data from a large selective research university to analyze the role of peers and grades in determining major persistence in the life and physical sciences. In the physical sciences, analyses using within-course, across-time variation show that ex-ante measures of peer quality in a student's introductory courses has a lasting impact on the probability of persisting in the major. This peer effect exhibits important non-linearities such that weak students benefit from exposure to stronger peers while strong students are not dragged down by weaker peers. In both the physical and life sciences, I find evidence that students are "pulled away" by their high grades in non-science courses and "pushed out" by their low grades in

their major field.

The final essay examines the effect of undergraduate course letter grades on future course selection and major choice. Using a Regression-Discontinuity design, I exploit the fact that the probability of earning a particular letter grade jumps discontinuously around letter grade cutoffs. This variation in letter grades allows me to isolate the impact of letter grades on major choice and course selection. I collect original numerical scores for 65 introductory courses across 6 fields and merge this with administrative data including student-level characteristics and transcripts. Since grading cutoffs exist throughout the distribution of scores, I am able to estimate local treatment effects at a variety of localities to examine the distribution of treatment effects. Contrary to the findings of the previous literature, I find no evidence that students respond to their letter grades in terms of course or major choices.

BIOGRAPHICAL SKETCH

Ben Ost (formerly known as Ben Wallenberg) was born in Jerusalem, Israel and grew up in Skokie, IL. He received his undergraduate degree in economics and math from Grinnell College and did his graduate work in Economics at Cornell University. Starting in the fall of 2011, he will be an assistant professor of economics at the University of Illinois at Chicago.

This dissertation is dedicated to Allie Rothschild. I am sorry that I did not previously give you the credit that you deserve.

ACKNOWLEDGEMENTS

I am deeply indebted to a number of individuals who made this dissertation not only possible, but pleasurable to write. First and foremost, I would like to thank Ron Ehrenberg for being constantly supportive and helpful over the past several years. From the moment I received his email inviting me to Cornell, Ron has without fail gone above and beyond and well out of his way to help me and my wife and for that I am truly grateful. My other committee members, Jordan Matsudaira and Kevin Hallock have provided countless helpful comments and suggestions and taught me most of what I know about econometrics and labor economics. I would also like to thank Matt Freedman, George Jakubson, Michael Lovenheim, and Emily Owens for many helpful conversations and comments on my research. I deeply appreciate the enormous help that Kirabo Jackson provided both in the form of comments and discussion and also through his invaluable course (which I took twice).

In addition to help from my committee, a number of other individuals provided useful comments and discussions that helped me learn economics and improve my dissertation. I would like to thank James Cowan, Catherine Maclean, Joyce Main, Mirinda Martin, Ross Milton, Eamon Molloy, Kristy Parkinson, Joshua Price, Kevin Roth, Michael Strain, Doug Webber and Ken Whelan for their beneficial peer effects.

The first chapter in my dissertation would not have been possible without the help of the North Carolina Education Research Data Center, and in particular Clara Muschkin and Kara Bonneau. The third chapter of the dissertation would not have been possible without the help of the numerous professors who provided grading data. Also, I am indebted to the entire Institutional Research and Planning Department at Cornell University, in particular Cathy Alvord,

Marin Clarkberg, Chari Fuerstenau, Dan McGough, Daniel J. Robertson, and Kevin Rexford for invaluable help with data used in my dissertation. I thank the Sloan Foundation and Cornell University for financial support.

I would like to thank my Saba and Savta for sending me snail mail articles relating to economics and education policy, and my brother Eyal, his partner Joe, my brother Noam and my in-laws Fred and Beth, for their love and support.

Penultimately, I would like to thank my parents, who think my research is silly but are still very supportive.

Finally, I would like to thank my wife Carrie Ost for everything. Half the dissertation was written by her (figuratively).

TABLE OF CONTENTS

INTRODUCTION

The three essays that comprise this dissertation are all based in the economics of education and examine issues of human capital acquisition.

In the first essay, I examine teacher productivity improvement and differentiate between general and specific human capital. Using micro-level longitudinal administrative data, I am able to construct exact task assignment histories for every teacher in North Carolina and I use this information to determine a teacher's previous experience, both at her current grade level and at other grade levels. By examining how general and specific experience help a teacher improve, I provide direct productivity based evidence of the relative importance of general and specific human capital. This is a significant contribution to the labor economics literature. The labor literature has previously inferred the role of general and specific human capital from the assumption that wages and marginal productivity are instantaneously equal, which goes counter to theories of wage deferring contracts or models of monopsonistic competition. I find magnitudes that are similar to those found in the general labor literature despite looking at productivity of teachers rather than wages in the general labor force. In addition to contributing to a large literature in labor economics, this essay is among the first that explores how teachers improve. Previous research has established that teachers perform better as they gain experience, especially in the first several years. My essay shows that teachers must be developing both general teaching skills, such as classroom management, as well as skills that are specific to a grade level such as curriculum familiarity.

The second essay is related to the first in that it explores human capital acquisition, but rather than examine how workers develop human capital on the

job, this essay explores why students choose to develop human capital in the sciences. In particular, I explore the determinants of persisting in the sciences conditional on having entered this field. A large fraction of students enter a science field but do not complete a degree and this is a major concern to policy makers. Science dropout is seen as weakening our global competitiveness in addition to acting as a major obstacle in diversifying scientific fields along gender and racial lines. Using longitudinal administrative data, I find that life and physical sciences exhibit dramatically different patterns of dropout, but several common factors influence persistence in the two fields. Relative grades are the most important correlate with persistence, such that students with higher science grades are more likely to persist in their scientific field, whereas students with higher grades in other fields are more likely to drop the science major. This paper also explores the impact of peer quality by comparing students across cohorts. I find that students are less likely to persist in the sciences when they take their introductory courses with other students who have a low likelihood of persisting given their fixed characteristics.

The third and final essay is written with Joyce Main and more rigorously tests the relationship between letter grades and major persistence documented in the second essay. While the positive correlation between introductory letter grades and persistence is clearly documented in the second essay, in addition to previous research, this correlation may not be causal. Theoretically, a student who cares about his or her GPA would rationally respond to low introductory grades by avoiding a major; however, it is equally true that students who are most committed to a particular major may also be those that choose to work hardest in that major. These two theories are both possible explanations of the relationship between grades and major choices and the policy implications de-

pend critically on which story is correct. In order to separately identify these theories, I collected exact numerical scores and merged these scores to administrative data that include transcript information. In this way, I constructed a data set that includes both the letter grade a student received in addition to the exact numerical score that she earned. This dataset allows me to test whether student major probabilities jump discontinuously with each letter grade (as would be the case if higher letter grades cause students to persist in a major) or whether student major probabilities are a smoothly increasing function of ones numeric score. I find little evidence that letter grades themselves causally increase the probability of majoring in a subject.

Though each essay in the dissertation is a distinct entity, a common thread of human capital runs through all three essays. The latter two essays are intimately related as they address essentially the same question of major choice, albeit using different methodologies. The first essay on teacher improvement breaks down a productive process rather than examining individual choices and shows that productivity can be significantly influenced by the exact nature of the human capital developed.

Chapter 1: How Do Teachers Improve? The Relative Importance of Specific and General Human Capital

1.1 Introduction

The degree to which human capital acquisition is general or specific has been of central concern in the labor economics literature since Becker (1964). While early research was primarily concerned with the degree to which human capital is specific to a firm or industry (Kletzer, 1989; Topel, 1991; Carrington, 1993; Neal, 1995; Parent, 2000), several recent papers suggest that the most relevant specificity may be based on the occupation or the tasks performed (Poletaev and Robinson, 2008; Kambourov and Manovskii, 2009; Gathmann and Schönberg, 2010). While this literature has made great strides in understanding the degree to which human capital is transferable across industries, firms, occupation and tasks, this literature measures productivity implicitly by assuming that wages perfectly reflect productivity in every time period. This key assumption fails to hold in the presence of wage-deferring contracts, differential monopsony power or efficiency wages and thus, any of these phenomena could substantially bias estimates of the degree of human capital specificity.

This paper provides new evidence of the relative importance of general and task-specific human capital by examining productivity improvement among teachers. A large literature has established that teachers improve with experience, but no previous study on teacher improvement has made the distinction between general teaching human capital and human capital that is specific to a particular grade level. Using micro-level longitudinal data, I track teachers, their grade-level assignments and a direct measure of productivity over a 13-year period. Using these data, I estimate the productivity improvements made by teachers as they gain general and grade-specific experience. This analysis provides estimates of how the entire history of teacher task assignments interact to determine current productivity.

The literature on teacher improvement has developed a high level of rigor in recent years thanks to the availability of matched teacher-student panel data. Rather than relying on cross-sectional

estimates which are likely flawed due to survival bias, researchers have used within teacher variation to determine productivity improvements due to experience. (Rockoff, 2004; Hanushek et al., 2005; Aaronson et al., 2007; Clotfelter et al., 2007) While this branch of research can establish that a teacher's performance increases over time, it is not able to explain *how* they are improving. If policy makers hope to increase the rate of improvement or lengthen it's horizon, knowing simply that teachers improve is uninformative. Ideally, researchers interested in helping teachers improve would be able to identify "best practices" that inexperienced teachers fail to implement and gradually learn. With specific information on what it is that teachers learn, policy makers and professional development specialists might be able to provide training which emphasizes these aspects and allows teachers to develop more rapidly. This paper provides a step in that direction by identifying the extent to which teachers gain general teaching human capital such as classroom management or grade-specific human capital such as curriculum familiarity.

1.2 Overview

This paper is the first to document two stylized facts.

1. Teachers switch grade assignments frequently within a school such that over a quarter teach two different grades in their first two years.

2. Students who have a teacher with more grade-specific experience make larger test score gains than students who have similarly experienced teachers with less grade-specific experience.

Fact #1 is critical to the implementation of my analysis because it suggests that sufficient variation exists between general and grade-specific experience. Fact #2 is suggestive of the main result of this paper, namely that grade-specific experience is important to teacher improvement. While suggestive, Fact #2 should not be taken as conclusive evidence that grade-specific experience is beneficial to teacher productivity. The difference in productivity between teachers of equal general experience levels could be the result of teachers improving with specific experience (as I argue), but could also potentially reflect differences in grade assignment patterns across different types of

schools and teachers. My analysis distinguishes between these possibilities by carefully considering the source of identifying variation and implementing tests for endogenous movement. First, because of the possibility that unobserved teacher characteristics are correlated with grade-specific experience, my preferred specification controls for teacher-by-school fixed effects and thus the primary findings are based on comparing a teacher to herself while at the same school. Second, in order to test for endogenous movement, I examine whether changes in grade-specific experience are predicted by current performance. I find no evidence that changes in grade-specific experience are systematically related to a teachers current performance. The stylized fact regarding teacher improvement remains true when placed in a regression context and importantly is robust to using fundamentally different sources of identifying variation. In particular the results are robust to the inclusion of school fixed effects, teacher fixed effects, teacher-by-school fixed effects, or teacher-by-school-by-grade fixed effects. Each of these specifications uses a fundamentally different source of identifying variation (discussed in the Empirical Model section) and yet the results are remarkably stable across specifications.

While the methodology used in this paper is in many ways similar to research using matched worker-firm data, a few important differences are worth mentioning. First, the literature using worker-firm data generally measures productivity implicitly by assuming that wages equal productivity in every period. This study uses a direct measure of productivity and thus avoids concerns that wage deferring contracts or similar mechanisms create a divergence between productivity and wages. Second a major concern in previous research is the possibility that workers switch to firms where they are better matched and thus match quality may be correlated with acquiring specific experience. The frequency with which teachers leave and return to tasks in my data allows me to implement a novel strategy that directly controls for unobserved match quality. Furthermore, because teachers frequently change grade assignments, I am also able to control for both the act of switching itself as well as the number of switches each teacher has experienced. My preferred estimates are thus identified from the specific pattern and order in which grades are taught.

To address this question, I use administrative matched teacher-student data that follows each teacher in the North Carolina public school system from 1995-2007. This data includes the exact grade assignment for each teacher in each year, and since students are tested annually, the data is well suited to estimate value-added models to assess teacher productivity in each year. Results from a value-added model show that teachers improve with experience and the magnitude of these improvements depends upon the frequency with which they are able to apply both general and grade-specific human capital.[1] In my preferred specification which includes teacher-by-school fixed effects, grade-specific experience is found to be approximately 50% as important as general experience for benefiting student math scores. There is little evidence that grade-specific experience benefits student reading scores. One potential explanation for why grade-specific experience benefits math but not reading is that in North Carolina the reading objectives are constant for third through fifth grade whereas the math objectives change each year.

This study's contributions are threefold. First, it provides direct empirical evidence that within an occupation, task-specific human capital acquisition can significantly affect productivity. Second, it overcomes many of the econometric concerns in previous research by examining an occupation where task assignments are the norm rather than the exception. Lastly, the results of this paper provide a more nuanced understanding of how teachers improve and can guide policy regarding teacher grade assignments and professional development.

[1]Value-added models rely on there being no systematic student sorting according to the teacher characteristic of interest. While previous research such as Rothstein (2010) and Clotfelter et al. (2006) show evidence of students sorting into classrooms, I find little evidence that this sorting is systematically related to experience within a teacher. I explore these issues in detail in the Identification Tests section.

1.3 Literature

1.3.1 Task-Specific Human Capital

While there is a wealth of empirical literature devoted to examining the relative importance of firm-specific and general human capital, relatively little research has empirically analyzed the role of task-specific human capital. In a theoretical paper on task-specific human capital, Gibbons and Waldman (2004) explain that "some of the human capital an individual acquires on the job is specific to the tasks being performed, as opposed to being specific to the firm" (p. 203). They argue that task-specific human capital has broad applicability compared to firm- or industry-specific human capital, and task-specific human capital can theoretically explain phenomena such as cohort effects, job design, and promotion patterns. Unfortunately, relatively few empirical studies have examined the role of task-specific human capital, possibly because most longitudinal data lacks descriptions of job tasks.

In two recent papers, Gathmann and Schönberg (2010) and Poletaev and Robinson (2008) examine task-specific human capital by examining worker transitions across jobs with different task requirements. Both papers create metrics to determine the distance between jobs in terms of the skills required. Using their metric, Gathmann and Schönberg (2010) find that task-specific human capital can account for up to 52% of wage growth. Similarly, Poletaev and Robinson (2008) finds that previous research showing that human capital is specific to the industry (Neal, 1995) is largely attributable to the return to specific skills.

In both Gathmann and Schönberg (2010) and Poletaev and Robinson (2008), the essence of the identification is comparing individuals who move to a job with a new set of task requirements to individuals who move to a job with similar task requirements as their previous employment. Though both papers focus on plant closings to plausibly assume exogenous displacements, both sets of authors note that there is still the possibility that unobserved differences exist between workers who move to a similar occupation (in terms of skills) and workers who move to a dissimilar occupation. Gathmann and Schönberg (2010) attempts to address this possibility by using local labor market

conditions as an instrument for job changes, while Poletaev and Robinson (2008) simply note that their estimates are descriptive and should not be interpreted causally. An advantage of both Gathmann and Schönberg (2010) and Poletaev and Robinson (2008) over my study is that the authors examine displaced workers from many occupations whereas my study is limited to teachers.

Another study analyzes the role of task-specific familiarity in the context of financial analyst forecasting performance (Clement et al., 2007). This paper uses cross-sectional analysis and tests whether analysts with more experience analyzing firm restructurings are more accurate in analyzing future restructurings than are other analysts. Their paper has the advantage of examining a narrowly defined type of specific experience and using a direct measure of productivity; however, because the data used is cross-sectional, unobserved heterogeneity across analysts cannot be ruled out. Importantly, the results found in this study may be driven by differential attrition among analysts.[2]

My paper is similar to Clement et al. (2007) in that my analysis is restricted to a single occupation and a narrowly defined "specific experience." Also, I observe a measure of productivity rather than assuming wages are equal to marginal product. Given the theoretical justification for a divergence between wages and marginal products over the life course (Lazear, 1979), observing productivity is a more direct test of the role of specific human capital. Unlike Clement et al. (2007), however, my data is longitudinal and thus I am able to control for unobserved time-invariant heterogeneity.

1.3.2 Teacher Experience

While the impact of many teacher characteristics is still debated, there exists an emerging consensus that teacher experience positively contributes to student learning, particularly for younger grades. Using data on middle school and elementary school students in Texas, Hanushek et al. (2005) find that students perform relatively worse when their teacher has less than three years of

[2]If analysts that perform poorly on their first firm-restructuring assignment are fired, then currently employed analysts with previous experience are likely better on average than currently employed analysts with no firm-restructuring experience.

experience. Rockoff (2004) finds consistent results using matched teacher-student data from two New Jersey elementary school districts. Similarly, Clotfelter et al. (2007) and Jackson and Bruegmann (2009) use the same North Carolina matched teacher-student data used in this paper and find that elementary teachers improve with experience, especially in the first several years.

The one exception to this consensus is Aaronson et al. (2007). The authors use data for ninth graders in the Chicago Public Schools and find no evidence of teachers improving with experience. One potential explanation for why the Aaronson et al. (2007) results differ from other studies is that most previous research has focused on students in grades 3-8 whereas Aaronson et al. (2007) considers high school teachers. For elementary grades, the fact that teachers typically teach the same students all day makes it more likely that differences in teaching ability will be detectable through student performance. Second, it is possible that the key skills that teachers develop as they gain experience are useful for teaching younger students but not for secondary education. Aaronson et al. (2007) is a clear exception to the literature given that in a meta-analysis of the value-added literature, Harris (forthcoming) finds that eight of nine studies show evidence of teachers improving with experience.

While many papers have demonstrated that teachers improve, the only paper of which I am aware that explores a mechanism for *how* teachers' on-the-job experience helps them improve is Jackson and Bruegmann (2009). The authors show that teachers improve when exposed to higher quality peers, thus demonstrating that part of teacher improvement is based on learning from other teachers. My paper builds on this research by identifying the type of skills that are most important to learn.

1.4 Data

I use longitudinal administrative data that links students to their teachers in the state of North Carolina between 1995-2007.[3] This data includes detailed information on student, classroom,

[3]This data has been extensively cleaned and standardized by the North Carolina Education Research Data Center housed at Duke University.

teacher, and school characteristics as well as a standardized measure of math and reading achievement for students in grades three through eight. For each student, the data include race, gender, parental education, free or reduced lunch status, and test scores for each grade. Available teacher characteristics include gender, race, highest degree earned, years of teaching experience, undergraduate institution, and licensure test scores.[4] Years of teaching experience is based on the number of years credited to a teacher for the purposes of salary calculation and thus should reflect all experience in any district.

By matching teacher information to classroom records, I am able to identify the grade taught by each teacher in each year. Using this information I construct a variable indicating the number of years a teacher has previously taught her current grade assignment. Because middle and high school teachers often teach multiple grades simultaneously, and because student test score data is most complete for third through fifth grade, I restrict my sample to elementary teachers who teach self-contained single-grade classes.

While the North Carolina data includes a link between student test scores and teachers, the teacher listed is actually the proctor of the student exam, not necessarily the classroom teacher. For elementary classrooms, the proctor is likely to be the classroom teacher, but to improve the accuracy of teacher-student matches, I limit the sample to confidently matched students. Following Clotfelter et al. (2007), Rothstein (2010), and Jackson and Bruegmann (2009) I consider a proctor to be the classroom teacher so long as the teacher's grade assignment matches the grade of the proctored exam. In addition, I drop cases where a proctor administered more than half of his/her tests to a different grade level.[5]

[4]This data is described in great detail in Clotfelter et al. (2007).

[5]In describing this data, Jackson and Bruegmann (2009) note that "According to state regulation, the tests must be administered by a teacher, principal, or guidance counselor. Discussions with education officials in North Carolina indicate that tests are always administered by the students' own teachers when these teachers are present. Also, all students in the same grade take the exam at the same time; thus, any teacher teaching a given subject in a given grade will almost certainly be administering the exam only to her own students."

To ensure comparability across years, I standardize all test scores by grade and year.[6] In order to implement some econometric specifications, I require a lagged test score in addition to current test scores. Students who are only present in the data for a single year are therefore dropped. The exception is for third graders, since the lagged third grade test is actually given to students at the beginning of the third grade rather than in second grade. While the data includes complete teacher data starting from 1995, complete student data is only available starting in 1997. I use the 1995-1996 period to calculate grade-specific experience, but only use 1997-2007 in regressions.

As discussed by Koedel and Betts (2008), achievement tests that contain ceilings may lead to systematic measurement error since students near the ceiling are unable to make further gains. In results not shown, I test for a ceiling in the North Carolina data by comparing a kernel density of each distribution to that of the normal density and find no cause for concern.

1.4.1 Data Limitation: Grade-Specific Experience

The data include information on teaching experience accrued before the sample period; however, my measure of grade-specific experience is limited to the sample time frame. For example, a teacher with ten years experience in 2003 accrued the latter eight years during the sample frame, but the data provides no information regarding the grades taught in her first two years (1993-1994). Thus, I cannot exactly determine this teacher's grade-specific experience for any year. In general, I cannot exactly calculate grade-specific experience for teachers who have pre-sample experience.

I address this data limitation with two distinct approaches. The first approach simply restricts the data to teachers who began teaching during the sample period and whose grade assignments are thus fully observed. The second approach uses all teachers for analysis and imputes grade-specific experience accrued prior to the sample period. This imputation is implemented by assuming that the histogram of grades a teacher taught out-of-sample matches the average histogram of grades taught by other teachers in her school. Each approach has benefits and drawbacks. Restricting

[6] Standardization is made prior to any data restrictions and thus the mean and standard deviation of the analysis sample is not exactly zero and one respectively.

the sample to fully observed teachers leads to an unrepresentitively inexperienced sample, while imputing grade-specific experience leads to measurement error in an explanatory variable. Because the measurement error introduced is not necessarily classical, however, the measurement error may bias estimates and it is difficult to assess its direction and magnitude.

Given that imputation may lead to biased estimates, my preferred specification restricts the sample to fully observed teachers though results are robust across both approaches. The fact that this sample of teachers is unusually inexperienced is likely a minor issue since previous research has found that most improvement occurs during the first several years (Rivkin et al., 2005).

Table 1.1 shows descriptive statistics for both the full sample and the sample which is restricted to fully observed teachers. As can be seen from this table, the restricted sample has considerably less experience on average than the full sample.[7] In addition, 25.6% of the full sample of teachers have an advanced degree whereas only 13.1% of the restricted sample of teachers have an advanced degree. The upper panel of Table 1.1 shows that restricting the sample to relatively inexperienced teachers also leads to a slightly different sample of students. Students in the restricted sample perform worse than students in the full sample and these students are also more likely to be a minority. These differences reflect that fact that schools with weaker, minority students, have relatively high teacher turnover rates and thus are disproportionately staffed by recently hired teachers. While the restricted sample of teachers is clearly not representative of teachers as a whole, it is the complete universe of recently hired teachers in the state of North Carolina and thus interesting in and of itself. The next section provides a more nuanced description of the relationship between grade-specific experience and general experience in this sample.

1.5 General Experience vs Grade-Specific Experience

In the absence of grade assignment changes, experience and grade-specific experience would be perfectly collinear and I would only be able to identify a single effect. To investigate the prevalence

[7] By definition, teachers in the restricted sample must have less than 13 years of experience whereas the full sample includes many teachers with 30+ years of experience.

of grade assignment changes, Table 1.2 presents a transition matrix showing grade assignments in year $t+1$ as a function of grade assignment in year t. This table demonstrates that over 25% of teachers switch grade assignments after teaching third, fourth or fifth grade. This table also documents that teachers are much more likely to switch to adjacent grades than distant grades. Evaluating whether experience in adjacent grades is more beneficial than experience in distant grades is hindered by the fact that relatively few teachers acquire experience in distant grades.

The frequent switching documented in Table 1.2 leads to a substantial divergence between experience and grade-specific experience. Table 1.3 presents a cross tabulation of grade-specific experience and experience for teachers in their first school. This table is restricted to teachers who have not switched schools to demonstrate that the divergence between general and grade-specific experience is not driven by school switching. Approximately 26% of teachers teach a new grade in their second year of teaching, and less than a half teach the same grade five times in their first five years teaching. This pattern continues in later years and suggests that it is possible to separately identify grade-specific and general experience for this sample.

In light of the fact that this paper shows that some human capital is specific to the task, it is somewhat surprising that so many switches occur. Conversations with principals indicate that in addition to the costs of grade switching documented in this paper, several benefits to task switching exist which may explain why teachers switch task assignments so frequently. First, switching teachers allows for a flexibility in management that can be useful when teacher teams conflict, lack diversity, or are uniformly experienced or inexperienced. Second, teachers who intend to become administrators may eventually benefit from the breadth of experience which comes from teaching a variety of courses. Third, it is possible that teachers grow bored with repetition over time and although they are better able to improve student test scores, their enthusiasm for the subject may wain. Finally, several principals indicated that they switch teachers to facilitate "professional growth history" — helping teachers become well rounded by having them teach a variety of grades.

No empirical estimates of the benefits to switching are available from the literature since no previous research has considered either the benefits or costs of teacher movement across grades

within a school. Although this paper presents evidence that some of teacher human capital is specific to a grade, the policy decision of whether to switch teachers must consider both the costs and benefits of switching. A complete cost-benefit analysis is beyond the scope of this paper, however, in the Identification Tests section, I examine whether patterns of teacher switching may be systematic in a way that would bias estimates of the return to specific experience.

1.5.1 Grade-Specific Experience and Student Performance

As a preliminary analysis of the effect of grade-specific experience, I perform simple mean comparisons of average student performance. Figure 1.1 graphically shows changes in average student test score gains as grade-specific experience varies. Each panel in this figure holds absolute years of experience constant and graphs the average student test score gains for teachers with various levels of grade-specific experience. These figures show that teachers with more grade-specific experience perform better in terms of their students' test score gains.

This relationship is especially clear for lower experience levels and is more pronounced for math score gains than reading score gains. While these figures are suggestive, they simply reflect raw correlations and by themselves cannot be interpreted as implying a causal relationship. However, in a later section I obtain a more controlled estimate of the impact of grade-specific experience which confirms the implications of the simple average comparisons.

1.6 Empirical Model

To evaluate the impact of teacher characteristics on student outcomes I use a value-added model (VAM) that controls for student characteristics, teacher characteristics, and several fixed effects to predict future test scores. My preferred specification controls for the lag of test score; however, I

explore the robustness of results across other value-added models in the appendix.[8]

$$A_{ijgst} = \alpha A_{ij_{t-1}g_{t-1}s_{t-1},t-1} + \beta X_i + \delta C_{ijgst} + \rho V_{ij} + \pi D_{jt} + f(Exp_{jt}) + g(Expgrd_{jt}) + S_{jt} + S_{jt}\ x\ f(Exp_{jt}) + S_{jt}\ x\ g(Expgrd_{jt}) + \xi_g + \omega_{js} + \phi_t + \varepsilon_{ijgst} \quad (1)$$

A_{ijgst} is the test score of student i taught by teacher j in grade g in school s in time t. The student characteristic vector X_i includes student gender, ethnicity, subsidized lunch status, and parental education. Classroom characteristics such as class size and average peer characteristics (excluding student i) are denoted by C_{ijgst}. The vector V_{ij} includes interactions between the student and teacher ethnicity and sex.[9] The vector D_{jt} includes a control for whether the teacher switched in the previous period and the total number of past switches a teacher has experienced. This model includes grade, teacher-by-school and year fixed effects denoted by ξ_g, ω_{js} and ϕ_t respectively. Experience and grade-specific experience enter through $f(\cdot)$ and $g(\cdot)$.

Since teacher performance may be disrupted by school switches, I include a series of dummy variables (S_{jt}) indicating whether the teacher is in her first, second or third school. Also, since the benefits to general or grade-specific experience may be specific to a school, I interact each of these indicators with general and specific experience in order to allow a structural shift when teachers change schools. The coefficients on the interaction of experience with whether a teacher is in her second or third school provides an estimate of the extent to which experience benefits are transferable across schools; however, I am hesitant to place much weight on these estimates because of the strong possibility that school changes, like any job change, are endogenous or reflect positive matching. Regardless, Appendix Table 1.12 shows that there is little evidence that the returns to experience change as teachers switch schools, suggesting there is a limited role for school-specific teaching skills. In other analyses (not shown) I include a cubic in school-specific

[8]The lagged test score VAM is used in many recent studies including Aaronson et al. (2007), Kane et al. (2006), Jackson and Bruegmann (2009), and others. Controlling for the lag of test score is found to outperform other value-added methodologies in an experimental validation study by Kane and Staiger (2008). In their paper Kain and Staiger find no evidence that non-experimental value-added estimates are biased, relative to experimental estimates.

[9]Dee (2005) shows that gender and ethnicity match may effect student achievement.

experience directly into the model and find no evidence that human capital is school specific. Given the possibility that school changes are endogenous, all estimates in this paper include non-parametric controls for school changes rather than relying on correctly estimating school-specific experience.[10]

As has been noted in previous research, measurement error in lagged test scores can bias estimates on all coefficients. I follow the procedure suggested by Anderson and Hsiao (1981) and Todd and Wolpin (2003) and use the second lagged test score as an instrument for the first lagged test score. Generally, this IV specification would drop any student who lacks three consecutive test scores leading to an unrepresentative sample that disproportionately represents students in relatively stable situations. I use the estimator proposed by Jackson and Bruegmann (2009) which avoids this significant data restriction. Essentially the Jackson and Bruegmann estimator uses the restricted student sample to estimate α using the double lag as an instrument for the lag of test score. The estimate of α is then used in estimating equation (1) for the entire sample.[11]

While Rothstein (2010) demonstrates significant non-random sorting of students into classrooms, this will only bias my results to the extent that this sorting is correlated with grade-specific teacher experience within a teacher. Furthermore, because I control for absolute years of teaching experience, estimates of the impact of grade-specific experience will only be biased if students are sorted into classrooms depending on the teachers' grade-specific experience conditional on a fixed level of overall teaching experience. I explore these concerns in the falsification section and find little evidence that within a teacher, students are differentially sorted as the teacher gains experience or grade-specific experience.

[10] Another possible approach would simply restrict the analysis to teachers in their first school, thus avoiding the possibility of endogenous movement across schools. This approach creates a sample selection issue however since teachers who never switch schools are likely different from the general sample. In practice, this alternative approach yields very similar estimates to those of the preferred model.

[11] See the online appendix of Jackson and Bruegmann (2009) for a proof of the consistency of this estimator.

Following Rockoff (2004), Aaronson et al. (2007), Koedel and Betts (2007), and others, I model experience effects as a cubic polynomial. Given the nearly perfect collinearity between general experience and year effects when teacher fixed effects are included, it is necessary to assume that teacher experience has no impact on quality after a certain threshold of experience. As in Rockoff (2004), Harris and Sass (2007), and Koedel and Betts (2007) I use 10 years as the cutoff for general experience, however results are not sensitive to either polynomial choice or the exact cutoff used.[12]

In order to test the sensitivity of results across various sources of identifying variation, I estimate equation (1) with teacher-by-school fixed effects, school fixed effects and teacher-by-school-by-grade fixed effects. To assess the relative merit of each of these identification strategies, it is useful to consider a decomposition of the error term from equation (1). For notational simplicity, I consider the error term from a model estimated at the teacher level; however, the intuition is identical for the full model. Consider estimating:

$$A_{jgt} = f(Exp_{jt}) + g(Expgrd_{jt}) + \beta X_j + \theta_{jgt} \tag{2}$$

In this regression, A_{jgt} is a productivity measure, Exp_{jt} is general experience, $Expgrd_{jt}$ is specific experience and X_j is a set of teacher characteristics. Without loss of generality, the error term from this regression, θ_{jgt}, can be broken down as follows (j denotes the teacher, g denotes a grade, and t denotes time):

$$\theta_{jgt} = \gamma_j + \gamma_g + \gamma_t + \nu_{jt} + \nu_{jg} + \nu_{gt} + \varepsilon_{jgt} \tag{3}$$

[12]See Rockoff (2004) for a discussion and justification of this assumption. In practice, this assumption is implemented by recoding the experience variable so that teachers with above ten years of experience are made to have exactly ten years of experience. In the unrestricted sample, this recoding leads to grade-specific experience exceeding general experience for the approximately 2% of teacher-year observations that have grade-specific experience above 10 years. By definition, experience must be greater or equal than grade-specific experience so I impose a cutoff of 10 years on grade-specific experience as well.

To the extent that any part of this error term is correlated with grade-specific experience, estimates from equation (2) will be biased. By including grade and time fixed effects γ_g and γ_t are completely absorbed and thus need not be considered. When only including school fixed effects however, there is reason to be concerned that γ_j might be correlated with experience. There are many mechanisms through which unobserved teacher quality may be correlated with grade teaching history within a school. First, it is possible that teachers generally prefer to repeatedly teach the same grade, and the best teachers are given this privilege. Conversely, it is possible that principals allow weaker teachers to repeatedly teach the same grade since higher quality teachers can more easily overcome the challenges of teaching a new curriculum. In addition to the possible endogeneity of grade switching, estimates of the effect of teacher experience may be biased if teacher exit rates are correlated with unobserved quality. If weaker teachers leave teaching relatively early in their career, then comparing experienced teachers to inexperienced teachers will lead to upwardly biased estimates of the return to experience.[13]

To address both the concerns of differential grade assignment and concerns of differential attrition, I include teacher-by-school fixed effects. With the inclusion of teacher-by-school fixed effects, grade fixed effects and time fixed effects, the error term θ_{jgt} simplifies to become:

$$\theta_{jgt} = \nu_{jt} + \nu_{jg} + \nu_{gt} + \varepsilon_{jgt} \tag{4}$$

To assess the likelihood that this error term is correlated with experience, it is useful to consider each piece separately. The first term (ν_{jt}) captures changes in productivity over time not captured by the controls for experience (functional form bias). So long as the return to experience is modeled in a flexible fashion, ν_{jt} is likely zero. The second term (ν_{jg}) captures the unobserved productivity match between teacher j and grade g. As predicted by a search model similar to Jovanovic (1979), this term is potentially correlated with grade-specific experience if teachers are more likely to stay teaching grades to which they are positively matched. The third term (ν_{gt}) captures global unobserved changes in grade difficulty over time. Since all measures of productivity are standardized

[13]Hanushek et al. (2005), Boyd et al. (2008) Goldhaber et al. (2007) all find evidence that less effective teachers are more likely to leave a school.

by grade-year, ν_{gt} is likely zero. The final term (ε_{jgt}) captures changes in unobserved productivity within a worker-task match. To the extent that unobserved teacher-grade match quality change over time, this can potentially bias estimates since principals may respond to match quality changes by switching teachers.

Most papers assume that ε_{jgt} is iid and are primarily concerned with addressing potential biases caused by ν_{jg}. In order to address this concern, many studies have focused on plant closings which lead to a job change plausibly unrelated to an individuals' current performance (Kletzer, 1989; Carrington, 1993; Neal, 1995; Gathmann and Schönberg, 2010; Kambourov and Manovskii, 2009). Furthermore, some studies have used local labor market conditions as an instrument to address endogenous mobility towards positive matches. While these methodologies potentially address endogenous mobility, they rely heavily on the validity of the instruments used or make other assumptions.

The frequency with which teachers leave and return to grades in my data allows me to implement a novel strategy that directly controls for unobserved match quality. My approach includes a teacher-by-grade fixed effect and thus the term ν_{jg} is completely absorbed and disappears from the error term. This approach is untenable in most analyses because within a worker-task, every time general experience increases, specific experience increases by the same amount. Thus, specific and general experience are perfectly colinear within a worker-task, even for individuals who switch jobs several times. The one exception is that when an individual leaves a task and later returns to it, in the interim period, the worker gains general experience but not specific experience and thus within a worker-task, the model is identified. Since my data includes a small number of potential tasks, a large number of workers and frequent switches between tasks, the data is well suited to estimating equation (2) with the inclusion of worker-by-task fixed effects.

With the inclusion of teacher-by-task fixed effects, ε_{jgt} remains a part of the error term and thus the key identifying assumption for this model is that to the extent that teachers are naturally better matched to certain grades, this natural match quality cannot change over time in a way that is systematically related to experience. Although including teacher-by-grade fixed effects effectively

addresses biases caused by match quality, one limitation of this approach is that it is identified primarily from teachers who switch grades and later return to the same grade.

1.7 Results

Results from estimating equation (1) are shown in Table 1.4 and Table 1.8. Consistent with previous research, a teacher's experience is found to positively impact student outcomes. While controlling for number of years of teaching (general experience), grade-specific experience also has a positive impact on student math scores.

Based on the restricted sample with teacher-by-school fixed effects, estimates for math scores imply that a teacher who teaches the same grade for each of her first five years helps students perform 0.140 standard deviations better than students with a novice teacher. If, instead, a teacher teaches different grades every year for her first five years, she helps students perform 0.0915 standard deviations better than a novice teacher. In other words, a fifth-year teacher who always repeats grade-assignments improves 52% more than a fifth-year teacher who never repeats grade assignments.[14] These magnitudes show fairly substantial improvement relative to the overall distribution of teacher quality and are consistent with previous estimates of the return to experience. To put these magnitudes in perspective, the difference between a novice teacher and a fifth-year teacher is similar to the impact found for policy interventions such as large class size reductions in Project STAR or movement to high performing charter schools. (Hoxby and Murarka, 2009; Schanzenbach, 2007) Estimates based on the full sample are similar to those from the restricted sample, and are shown in the appendix.

Based on the restricted sample there is no statistically significant effect of grade-specific experience on reading scores. Estimates based on the full sample, however, show small statistically significant benefits of grade-specific experience for reading scores. Regardless, the magnitude of specific effects for reading is considerably smaller than for math and given the lack of robust-

[14]While a fifth-year teacher is considered for expository purposes, the benefits to grade specific experience are roughly fifty percent throughout the improvement profile

ness across data choices, I am hesitant to conclude that grade-specific experience benefits reading scores. While I have no definitive explanation as to why grade-specific experience matters more for math than for reading, one possible cause is the fact that similar reading skills are taught in each grade whereas math curricula change dramatically for each grade. The North Carolina standard curriculum "five competency goals" demonstrate this point. Between third and fifth grades, all five reading competency goals remain identical for each grade whereas all five math competency goals change for each grade (North Carolina Department of Education, 2009).

The second and fifth columns of Table 1.4 show that results are similar when using school fixed effects. The model with school fixed effects has the advantage of utilizing a larger amount of variation to identify the grade-specific experience effect, since teachers who never switch grades still contribute to these estimates. However, given the multiple avenues through which school fixed effects models could be biased, these are not my preferred specifications and are shown only as a robustness check. Despite several potential avenues for bias, the magnitudes found when only including school fixed effects are very similar to those when including teacher-by-school fixed effects. A teacher who teaches the same grade for each of her first five years helps students perform 0.146 standard deviations better than a novice teacher. If instead a teacher teaches different grades every year for her first five years, she helps students perform 0.086 standard deviations better than a novice teacher. These estimates are statistically indistinguishable from the estimates that include teacher-by-school fixed effects.

Models with teacher-by-school-by-grade fixed effects have the advantage of controlling for unobserved differences in a teacher's inherent ability to teach different grades; however, these estimates are identified only for teachers who switch grades and then switch back to their original grade. Since only 19.11% of teachers switch grades and then switch back, estimates that include teacher-by-school-by-grade fixed effects are identified from a relatively small fraction of teachers. Estimates that include teacher-by-school-by-grade fixed effects are fairly similar to estimates that only include teacher-by-school fixed effects suggesting that grade-matching is a minor concern. Based on estimates that include teacher-by-school-by-grade fixed effects, a teacher who teaches

the same grade for each of her first five years helps students perform 0.168 standard deviations better than students with a novice teacher. If instead a teacher teaches different grades every year for her first five years, she helps students perform 0.091 standard deviations better than a novice teacher. These estimates are statistically indistinguishable from the estimates that only include teacher-by-school fixed effects.

Given that results are similar when including school, teacher-by-school, or teacher-by-school-by-grade fixed effects, it appears that the exact source of identifying variation is relatively unimportant. That said, my preferred specification includes teacher-by-school fixed effects because it controls for possibly important unobserved heterogeneity without severely restricting the identifying sample.

1.7.1 Experience in Adjacent Grades

To the extent that age-specific teaching skills are important, experience in nearby grades may be more relevant than experience in distant grades. To examine this possibility, I calculate a measure of "nearby" experience; specifically, for each teacher, I count the number of years in which that teacher has taught either her current grade or an adjacent grade. This specification thus distinguishes between three types of experience: general experience, grade-specific experience, and grade-or-adjacent-grade experience. Controlling for general and grade-specific experience, the impact of grade-or-adjacent-grade experience gives the benefit of nearby experience above and beyond the benefit of general experience. For consistency, I model each type of experience as a cubic and include the same controls as found in the primary analysis.

Table 1.5 shows the results for this regression when using various fixed effects. While the point estimates from some specifications suggest that nearby grades benefit performance above and beyond general experience, the results are not statistically significant in most specifications. Furthermore, the coefficient for grade-specific experience ceases to be significant in my preferred specification with teacher-by-school fixed effects. These results are not surprising given that Table 1.2 shows that teachers typically move only between adjacent grades and thus general and grade-

or-adjacent-grade experience are highly correlated ($\rho = 0.89$). In addition, the estimates shown in Table 1.5 are fairly sensitive to specification checks, providing further concern that collinearity may be an issue. Rather than concluding that all 3 types of experience are not important determinants of teacher performance, I view these results as indicating that too little variation exists between these three measures of experience to be able to separately identify effects.

1.8 Identification Tests

When using teacher fixed effects, the effect of grade-specific experience is identified by two different sources of variation. First, grade-specific experience diverges from general experience whenever a teacher receives a new grade assignment. Second, conditional on the number of times a teacher switches grades, grade-specific experience varies depending on the pattern of grades taught and the order in which these grades are taught.

1.8.1 Test for Systematic Grade Switching

Since one source of variation is based on grade switching, estimates which use this source of variation may be biased if teacher switching is correlated with expected performance.[15] If teachers are switched to grades in which they have less experience in years when one expects that they will do particularly poorly, this may lead to overstating the importance of grade-specific experience. To test whether teacher switching is related to expected performance, I estimate equations (5) and (6). These equations test whether teachers are switched based on current performance. Since it is impossible to directly measure "expected" performance, I test whether current performance predicts whether a teacher is switched the following year. Since principals might consider raw test scores in addition to test score progress, I define performance as absolute test score in equation (5) and as test score gains in equation (6). I use a linear probability model (LPM) to predict whether a teacher switches grade assignments between years. This model is run as a linear probability model

[15] More exactly, estimates may be biased if teacher switching is correlated with expected performance conditional on all observables.

rather than a non-linear model for simplicity and because empirically, predicted probabilities all lie between zero and one using the LPM.

$$\mathbf{1}(g_{j,t} = g_{j,t+1}) = \zeta \bar{A}_{j_t g_t s_t t} + \lambda \sum_g \left| N_{gs(t+1)} - N_{gs(t)} \right| + \beta E_{jt} + \omega_{js} + \xi_g + \phi_t + \varepsilon_{jt} \quad (5)$$

$$\mathbf{1}(g_{j,t} = g_{j,t+1}) = \zeta \overline{\Delta A}_{j_t g_t s_t t} + \lambda \sum_g \left| N_{gs(t+1)} - N_{gs(t)} \right| + \beta E_{jt} + \omega_{js} + \xi_g + \phi_t + \varepsilon_{jt} \quad (6)$$

The variable $\mathbf{1}(g_{j,t} = g_{j,t+1})$ is an indicator that is unity when teacher j repeats grade assignments and zero when teacher j switches grade assignments. The variable $\bar{A}_{jgst}$ denotes the average test scores for students taught by teacher j in grade g in school s in time t.[16] The vector E_{jt} includes experience and grade-specific experience in period t. Both specifications include teacher-by-school, grade, and year fixed effects denoted by ω_{js}, ξ_g and ϕ_t respectively. The variable N_{gst} is the number of sections of grade g in year t and thus the coefficient λ captures the impact of school-by-year changes in the demand for teachers of each grade.[17] In performing this test, I am primarily interested in the coefficient ζ because this coefficient reflects the extent to which teachers are switched due to their current students' performance.[18]

As can be seen in the first two columns of Table 1.6, there is little evidence that teachers are switched based on current performance. A one standard deviation improvement in average reading

[16]These regressions are run at the classroom level because the tests aim to capture principals' responses to class performance. When the same regressions are estimated at the student level, the results are similar but the magnitudes of the estimates are considerably smaller.

[17]Hoxby (2000) exploits population variation to identify the effect of class size on student achievement. Similarly, population variation leads to changes in the number of classes per grade. When a particularly large cohort of students passes through a school, teachers may need to be switched around each year in order to create extra sections for the large cohort. The expression $\sum_g \left| N_{gs(t+1)} - N_{gs(t)} \right|$ gives the total number of section changes in a school for a given year. I include this variable because as the number of section changes increase, I expect that teachers are more likely to switch grades to fill those empty positions.

[18]If one teacher switches grades, another teacher will possibly need to switch grades as well, thus standard errors are clustered at the school-year level. In practice, using OLS standard errors or other clustering levels lead to similar results.

test gains leads to a 0.9 percentage point decrease in the probability of being switched. There is a small negative effect from math gains as well; however this is not statistically significant. The level of average reading scores and math scores are not statistically significant predictors of teacher switching. Because there is some evidence that teacher switches are weakly correlated with performance, my empirical model controls both for switching last period and the total number of switches a teacher has experienced.

1.8.2 Test for Systematic Changes in Grade-Specific Experience

The second (and more exact) source of identifying variation is based upon the magnitude of grade-specific experience changes. For example, when a teacher switches grade assignments, this generally leads to a decrease in grade-specific experience; however, grade-specific experience sometimes increases as a result of a switch.[19] When controlling for both whether a teacher switches and current grade-specific experience, estimates will be biased if next year's grade-specific experience is correlated with expected performance. I test for this correlation by examining whether current teacher performance predicts changes in next year's grade-specific experience. It should be emphasized that this test cannot detect correlations between grade-specific experience and expected performance that are unrelated to current performance. I estimate:

$$\Delta EXPGRD_{j,t+1} = \gamma \bar{A}_{j_t g_t s_t t} + \beta E_{jt} + \omega_{js} + \xi_g + \phi_t + \delta s_{jt} + \varepsilon_{jt} \tag{7}$$

$$\Delta EXPGRD_{j,t+1} = \gamma \overline{\Delta A}_{j_t g_t s_t t} + \beta E_{jt} + \omega_{js} + \xi_g + \phi_t + \delta s_{jt} + \varepsilon_{jt} \tag{8}$$

The variable $\Delta EXPGRD_{t+1}$ denotes the change in grade-specific experience for teacher j between time t and $t+1$, s_{jt} is an indicator for whether the teachers switched and all other variables are defined as in (5) and (6). As shown in the third and fourth column of Table 1.6 there is no evidence that grade-specific experience changes are related to current performance.

Based on the above tests, I conclude that conditional on switching, the resulting grade-specific experience variation is not systematically related to previous teacher performance.

[19] In my data, 13.45% of switches result in increases in grade-specific experience.

1.8.3 Test for Student Sorting

A major concern in using value-added models to measure teacher productivity is the extent to which students are non-randomly assigned to teachers. Importantly, the non-random assignment of students to teachers is necessary but not sufficient to bias estimates of experience effects. In order for non-random student assignment to bias experience effects, student assignment must be systematically related to teacher experience within a teacher. Just as random student assignment is sufficient for eliminating bias in the presence of non-random teacher placement, randomly assigning teachers (according to their experience) is sufficient for eliminating biased estimates of experience effects, in the presence of non-random student sorting.

Rothstein (2010) provides evidence that students are non-randomly assigned to teachers; however, since I include teacher-fixed effects, bias is only created if students are sorted differentially over a teacher's tenure. In addition, as demonstrated by Rothstein (2009), the lagged test score model that I use has the advantage that it can account for dynamic systematic sorting based on observable past student performance. While dynamic sorting based on unobserved performance can still bias estimates in a lagged model, specifications that include student-fixed effects and exclude the lagged score are unable to even control for observable dynamic student sorting.

With this type of bias in mind, I test for systematic assignment based on grade-specific experience. Following Jackson and Bruegmann (2009), I test for this type of bias by examining whether a future teacher's experience is correlated with current student outcomes, conditional on observables. While I confirm Rothstein's finding that a student's future teacher appears to have an "effect" on current performance, I find weak evidence that this non-random sorting is related to grade-specific experience.[20]

Table 1.7 shows the results of a regression in which the experience levels of each student's future teacher is included in the regression given by equation (1). This regression examines whether

[20]This indicates that my estimates of grade-specific experience effects are likely not affected by non-random sorting, but individual teacher fixed effects should not be relied upon to evaluate that specific teacher's effectiveness.

a student's future teacher's experience levels are correlated with a student's current performance, conditional on observables. As can be seen in this table, there is no evidence of non-random sorting based on teacher grade-specific experience.

Because of the nature of the falsification exercise, students must be observed in three consecutive years to be included in the regression. Furthermore, the test requires that the future teacher is fully observed in the sample (in order to calculate grade-specific experience) and thus drops any student who's future teacher has out-of-sample experience. This restriction leads to a considerable sample size reduction which lowers the power of these regressions. It is possible that estimates given in Table 1.7 would be statistically significant if the sample size was equivalent to Table 1.4. As shown in the appendix, when using the full sample of teachers (with grade-specific experience imputations), there is some evidence that students are sorted according to a teacher's general experience level for math scores. Regardless, the magnitudes of these false experience "effects" are quite small compared to the estimates given in Table 1.4. Based on this rather weak evidence of student sorting, I conclude that it is unlikely that previous estimates are substantially biased due to student sorting.

1.9 Alternative Interpretations

Based on the results shown above, I conclude that teachers likely benefit from both general and grade-specific experience, however a number of alternative interpretations are possible as well. First, it is possible that rather than benefitting from grade-specific experience, teachers suffer from disruption. All specifications directly control for both a disruption indicator as well as a count of the number of total disruptions, however, to the extent that disruption impacts long-term, but not short term outcomes and does so in a non-linear fashion, my controls for disruption effects may be insufficient. I find this interpretation a less likely explanation than a specific human capital explanation, but am not able to definitively rule it out.

Another possible interpretation of the results is that teachers do not improve in general, but simply become more familiar with the examinations and thus become more effective at preparing

students for the tests. This interpretation could explain all past research on teacher improvement and is difficult to disprove. If this interpretation of the results is correct however, it fundamentally alters the optimal policy recommendations. Regardless, given that currently principals and administrators aim to improve test scores, a teacher that is more effective at preparing students for a test might be considered more productive by the administration.

1.10 Conclusion

The fact that teachers improve with experience is commonly cited as one reason why teacher attrition is problematic. This paper shows that frequently reassigning a teacher to a new grade has consequences similar to teacher attrition because his or her grade-specific human capital is wasted. While it is very difficult and expensive to affect teacher attrition through policy, improving teacher grade assignments is more straightforward to implement. Based on conversations with principals and teachers, it is apparent that completely avoiding grade assignment switches is unrealistic. In cases where grade reassignments are unavoidable, however, principals should consider providing teachers who are new to their grade assignment many of the supports provided to teachers who are generally inexperienced.

More generally, the evidence presented in this paper suggests that the applicability of human capital is directly related to the similarity between past and future tasks performed. Thus, in analyzing the effect of worker movement, the most relevant question may not be whether the worker moves to the same industry or firm, but rather whether the new job requires similar tasks to the old. Gathmann and Schönberg (2010) provide some evidence of the importance of task-specific human capital in the context of job changes, but this literature remains under developed.

While the availability of panel data has been used extensively to make methodological improvements to previous research on teacher quality, relatively few studies have used the detailed longitudinal data to analyze in detail how the past impacts the future. This study measures not only whether a teacher was teaching five years earlier, but considers *what* she was teaching. Using this dynamic measure of task assignments within a school, this paper separately identifies the pro-

ductivity benefits of general and grade-specific human capital and finds that both are important in determining the rate of teacher improvement.

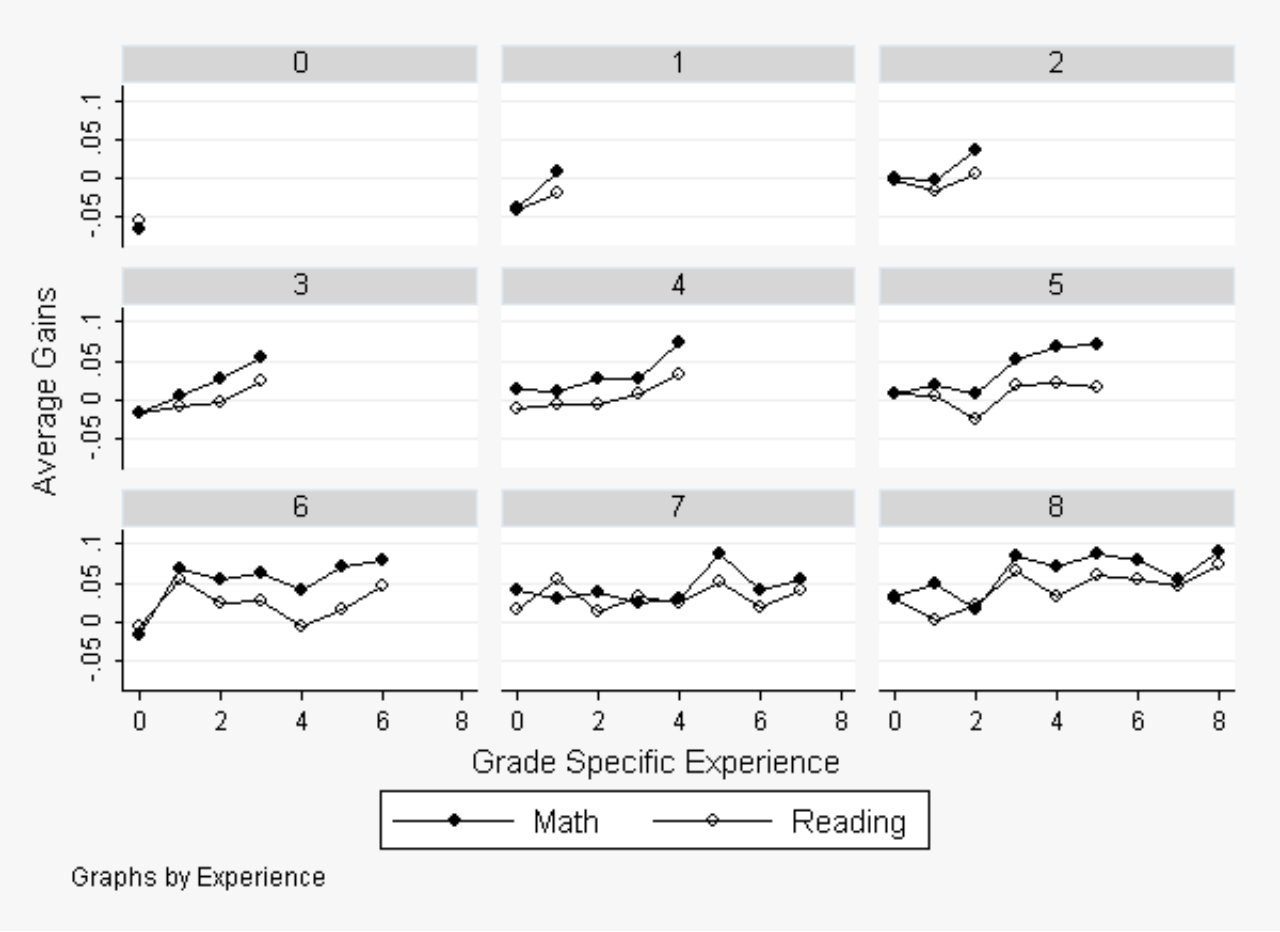


Figure 1.1: Average student test score gains by teacher grade specific experience: Split by experience level.

Notes: As experience rises, the number of teachers reflected in each point sharply decreases and thus the plots become increasingly noisy. For example, there are relatively few teachers with exactly 8 years of experience *and* exactly 6 years of grade-specific experience.

Table 1.1: Descriptive Statistics

	Full Sample			Restricted Sample		
Variable	Observations	Mean	Standard deviation	Observations	Mean	Standard deviation
Unit of observation: Student-year						
Math Score	2188251	0.041	0.99	688962	-0.02	0.984
Reading Score	2178575	0.037	0.988	685667	-0.03	0.988
Change in math score	1840340	0.026	0.704	562480	0.011	0.733
Change in reading score	1832654	0.014	0.732	560052	-0.006	0.764
Female	2189917	0.495	0.5	689539	0.495	0.5
Black	2189917	0.284	0.451	689539	0.302	0.459
Hispanic	2189917	0.055	0.228	689539	0.068	0.252
Parent is high school dropout	2189917	0.094	0.291	689539	0.082	0.274
Parent is high school graduate	2189917	0.560	0.496	689539	0.542	0.498
Parent is college graduate	2189917	0.236	0.425	689539	0.221	0.415
Class size	2189917	22.909	3.313	689539	22.525	3.362
Student has limited english proficiency	2189917	0.019	0.137	689539	0.022	0.146
Unit of observation: Teacher-year						
Experience	100681	12.051	9.585	32306	2.911	2.715
Grade-specific experience (No Imputation)	-	-	-	32306	1.842	2.109
Grade-specific experience (Imputed)	100681	5.685	3.935	-	-	-
Female Teacher	100681	0.928	0.259	32306	0.878	0.327
Black Teacher	100681	0.138	0.345	32306	0.113	0.317
Hispanic Teacher	100681	0.003	0.059	32306	0.005	0.068
Teacher has advanced degree	100681	0.256	0.436	32306	0.131	0.338
Teacher Exam Score	93993	0.039	0.835	29913	0.152	0.702

Table 1.2: Grade Assignment Transition Matrix

Grade taught in year t	Grade taught in year t+1									
	PK	K	1	2	3	4	5	6	7 and up	Total
PK	82.7%	10.6%	2.7%	1.9%	1.2%	0.3%	0.2%	0.1%	0.3%	100.0%
K	1.0%	78.6%	9.9%	3.7%	2.4%	1.4%	1.1%	0.9%	1.0%	100.0%
1	0.3%	9.6%	74.3%	8.2%	3.4%	1.6%	1.2%	0.7%	0.7%	100.0%
2	0.3%	4.6%	7.9%	72.0%	9.2%	3.0%	2.0%	0.6%	0.6%	100.0%
3	0.2%	3.4%	3.8%	6.8%	71.8%	8.1%	3.6%	1.4%	1.0%	100.0%
4	0.2%	2.4%	2.5%	3.8%	7.4%	72.1%	8.7%	1.7%	1.3%	100.0%
5	0.1%	1.8%	1.5%	2.2%	4.8%	8.1%	75.4%	4.0%	2.0%	100.0%

Note: This table is restricted to teachers who teach in the same school in year t and t+1 and are fully observed in the data.

Table 1.3: Grade Specific Experience by Total Experience

Experience	Grade Specific Experience						
	0	1	2	3	4	5	Total
0	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
1	26.3%	73.7%	0.0%	0.0%	0.0%	0.0%	100.0%
2	18.1%	19.2%	62.7%	0.0%	0.0%	0.0%	100.0%
3	14.8%	13.7%	16.4%	55.1%	0.0%	0.0%	100.0%
4	13.0%	10.5%	11.8%	14.7%	50.0%	0.0%	100.0%
5	11.6%	10.2%	8.5%	10.6%	14.7%	44.4%	100.0%

Note: This table includes teachers in the fully observed sample who have not switched schools.

Table 1.4: Impact of Teacher Experience on Student Performance

	Math			Reading		
Fixed Effects:	Teacher-by-School (1)	School (2)	Teacher-by-School-by-Grade (3)	Teacher-by-School (4)	School (5)	Teacher-by-School-by-Grade (6)
Grade-specific experience	0.0211*	0.0374**	0.0325†	-0.0013	0.0083	0.0075
	(0.0087)	(0.0082)	(0.0171)	(0.0083)	(0.0070)	(0.0166)
(Grade-specific experience)^2	-0.0033	-0.0071**	-0.0039	0.0017	-0.0010	0.0022
	(0.0024)	(0.0023)	(0.0026)	(0.0023)	(0.0020)	(0.0025)
(Grade-specific experience)^3	0.0002	0.0004†	0.0001	-0.0002	0.0000	-0.0002
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Experience	0.0453**	0.0347**	0.0436*	0.0624**	0.0321**	0.0647**
	(0.0129)	(0.0076)	(0.0190)	(0.0126)	(0.0064)	(0.0183)
Experience^2	-0.0079**	-0.0055**	-0.0076**	-0.0088**	-0.0058**	-0.0094**
	(0.0021)	(0.0020)	(0.0023)	(0.0020)	(0.0017)	(0.0022)
Experience^3	0.0005**	0.0004*	0.0005**	0.0006**	0.0004**	0.0007**
	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(0.0002)
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Grade fixed effects	Yes	Yes	No	Yes	Yes	No
Student characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Peer characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Observations	560874	560874	560874	558451	558451	558451

† Significant at 10%; * significant at 5%; ** significant at 1%. Standard errors clustered at class level reported in parentheses.

Notes: The dependent variable is a standardized measure of test score. Only teachers who begin teaching during the sample frame are included in this regression. Appendix Table 1.12 shows the complete set of coefficients from these regressions.

Table 1.5: Impact of Teacher Experience in Adjacent Grades on Student Performance

	Math			Reading		
Fixed Effects:	Teacher-by-School (1)	School (2)	Teacher-by-School-by-Grade (3)	Teacher-by-School (4)	School (5)	Teacher-by-School-by-Grade (6)
Grade-specific experience	0.0144	0.0204*	0.0265	-0.0037	-0.0015	0.0012
	(0.0096)	(0.0091)	(0.0197)	(0.0092)	(0.0078)	(0.0192)
(Grade-specific experience)^2	-0.0016	-0.0037	-0.0017	0.0021	0.0005	0.0034
	(0.0027)	(0.0026)	(0.0030)	(0.0027)	(0.0023)	(0.0030)
(Grade-specific experience)^3	0.0001	0.0001	0.0001	-0.0002	-0.0000	-0.0003
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Experience	0.0336*	0.0186*	0.0352	0.0513**	0.0216**	0.0495*
	(0.0142)	(0.0089)	(0.0224)	(0.0140)	(0.0074)	(0.0211)
Experience^2	-0.0056*	-0.0030	-0.0058*	-0.0076**	-0.0045*	-0.0084**
	(0.0025)	(0.0023)	(0.0027)	(0.0023)	(0.0019)	(0.0026)
Experience^3	0.0004*	0.0002	0.0005*	0.0006**	0.0004**	0.0006**
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0002)
(Exp. in grd or adjacent grd)	0.0167	0.0331**	0.0073	0.0060	0.0211**	0.0140
	(0.0103)	(0.0086)	(0.0214)	(0.0098)	(0.0070)	(0.0192)
(Exp. in grd or adjacent grd)^2	-0.0040	-0.0058*	-0.0043	-0.0018	-0.0032†	-0.0024
	(0.0025)	(0.0024)	(0.0028)	(0.0023)	(0.0019)	(0.0026)
(Exp. in grd or adjacent grd)^3	0.0002	0.0003†	0.0002	0.0001	0.0001	0.0001
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0002)
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Grade fixed effects	Yes	Yes	No	Yes	Yes	No
Student characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Peer characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Observations	560693	560693	560693	558268	558268	558268

† Significant at 10%; * significant at 5%; ** significant at 1%. Standard errors clustered at class level reported in parentheses.

Notes: The dependent variable is test score. Only teachers who begin teaching during the sample frame are included in this regression.

Table 1.6: Test for Dynamic Endogeneity of Grade-Specific Experience

	Restricted sample				Full sample			
	Prob. teacher is switched between year t and t+1		Δ Grade-specific exper. after year t		Prob. teacher is switched between year t and t+1		Δ Grade-specific exper. after year t	
Ave. math score in year t		-0.014		0.020		-0.005		0.008
		(0.010)		(0.031)		(0.004)		(0.016)
Ave. read score in year t		0.000		-0.007		-0.004		0.002
		(0.009)		(0.030)		(0.004)		(0.015)
Ave. math gains in year t	-0.003		-0.009		-0.000		0.001	
	(0.006)		(0.018)		(0.002)		(0.010)	
Ave. read gains in year t	-0.009†		-0.008		-0.005*		-0.006	
	(0.005)		(0.018)		(0.002)		(0.009)	
# course chngs t to t+1	0.000	0.001	0.001	-0.000	0.002†	0.002†	0.000	-0.001
	(0.002)	(0.002)	(0.006)	(0.006)	(0.001)	(0.001)	(0.003)	(0.003)
General exp. in year t	0.042†	0.041†	-0.036	-0.005	0.008**	0.007**	0.070**	0.068**
	(0.024)	(0.022)	(0.094)	(0.082)	(0.002)	(0.002)	(0.007)	(0.006)
Was switched t to t+1			-2.159**	-2.177**			-2.305**	-2.314**
			(0.083)	(0.084)			(0.054)	(0.053)
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grade fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effect:	Teacher-by-School	Teacher-by-School	Teacher-by-School	Teacher-by-School	Teacher-by-School	Teacher-by-School	Teacher-by-School	Teacher-by-School
Observations	18764	20179	18725	20138	66874	71192	66762	71076

† Significant at 10%; * significant at 5%; ** significant at 1%. Standard errors reported in parentheses

Notes: The observation is at the class level so test scores (and gains) are class averages. For the first two columns, the dependent variable is an indicator for whether a teacher switches grades between year t and year t+1. For the last two columns, the dependent variable is an indicator for whether a teacher switched grades between year t-1 and year t.

Table 1.7: Falsification: "Impact" of Future Teacher Experience on Current Student Performance

	Math			Reading		
Fixed Effects:	Teacher-by-School (1)	School (2)	Teacher-by-School-by-Grade (3)	Teacher-by-School (4)	School (5)	Teacher-by-School-by-Grade (6)
Lead teacher grade exper.	0.0037	-0.0040	0.0042	-0.0003	-0.0061	0.0008
	(0.0161)	(0.0160)	(0.0165)	(0.0172)	(0.0164)	(0.0177)
(Lead teacher grade exper.)^2	0.0016	0.0038	0.0018	0.0031	0.0049	0.0034
	(0.0046)	(0.0048)	(0.0047)	(0.0049)	(0.0049)	(0.0051)
(Lead teacher grade exper.)^3	-0.0002	-0.0003	-0.0002	-0.0003	-0.0004	-0.0003
	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)
Lead teacher exper.	-0.0001	0.0033	-0.0016	0.0029	0.0043	0.0007
	(0.0150)	(0.0146)	(0.0154)	(0.0162)	(0.0149)	(0.0167)
(Lead teacher exper.)^2	-0.0009	-0.0016	-0.0005	-0.0010	-0.0018	-0.0008
	(0.0040)	(0.0039)	(0.0041)	(0.0043)	(0.0040)	(0.0044)
(Lead teacher exper.)^3	0.0001	0.0001	0.0001	0.0000	0.0001	0.0000
	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
Observations	87959	87959	87959	87595	87595	87595

† Significant at 10%; * significant at 5%; ** significant at 1%. Standard errors clustered at class level reported in parentheses.

Notes: The dependent variable is test score in period t. Lead experience and lead grade-specific experience correspond to the characteristics of student i's teacher in period t+1. All period t covariates are included in this regression, so this test is equivalent to testing whether lead teacher characteristics are correlated with unobserved student performance.

A Robustness Across Value Added Models

The lagged IV model used in the paper is my preferred specification both because it has performed well in experimental validation studies and because it can control for student sorting so long as that sorting is based on past test score. For robustness, I estimate other common VAMs in this appendix. Using the same variable definitions as presented in the paper, three commonly estimated VAM models are laid out below.

Gains model

$$\begin{aligned} \Delta A_{ijgst} = \beta X_i + \delta C_{ijgst} + \rho V_{ij} + \pi D_{jt} + f(Exp_{jt}) + g(Expgrd_{jt}) \\ + S_{jt} + S_{jt}\ x\ f(Exp_{jt}) + S_{jt}\ x\ g(Expgrd_{jt}) + \xi_g + \omega_{js} + \phi_t + \varepsilon_{ijgst} \end{aligned} \tag{9}$$

Lagged test score model

$$\begin{aligned} A_{ijgst} = \alpha A_{ij_{t-1}g_{t-1}s_{t-1},t-1} + \beta X_i + \delta C_{ijgst} + \rho V_{ij} + \pi D_{jt} + f(Exp_{jt}) + g(Expgrd_{jt}) \\ + S_{jt} + S_{jt}\ x\ f(Exp_{jt}) + S_{jt}\ x\ g(Expgrd_{jt}) + \xi_g + \omega_{js} + \phi_t + \varepsilon_{ijgst} \end{aligned} \tag{10}$$

Student fixed-effect model

$$\begin{aligned} A_{ijgst} = \beta X_i + \delta C_{ijgst} + \rho V_{ij} + \pi D_{jt} + f(Exp_{jt}) + g(Expgrd_{jt}) \\ + S_{jt} + S_{jt}\ x\ f(Exp_{jt}) + S_{jt}\ x\ g(Expgrd_{jt}) + \mu_i + \omega_{js} + \phi_t + \varepsilon_{ijgst} \end{aligned} \tag{11}$$

The lagged test score model is identical to the preferred specification from the text except that it does not instrument for the lagged test score. This model assumes that all past inputs can be summarized by the lagged test score and it implicitly assumes that the effects of past inputs decay geometrically. The gains model additionally makes the assumption that $\alpha = 1$. One advantage of the gains model over the lagged model is that it completely avoids the measurement error problems of including the lagged score as an independent variable.

The student fixed effects model controls for unobserved fixed characteristics of students and assumes that past inputs have no effect on future outcomes. Unfortunately, given my data, the model with student fixed effects may be unreliable when estimated for the fully observed, restricted

sample of teachers. This analysis is complicated by the fact that when I restrict the sample to fully observed teachers, this severely reduces the number of students who are observed for two or more years.[21] As a result, while restricting the sample to fully observed teachers is broadly my preferred specification, since it destroys the continuous panel structure along the student dimension, my preferred student fixed effect estimates are based on the full sample of teachers.

Even when using the full sample of students however, one drawback to using the student fixed effects model is that one cannot simultaneously control for grade fixed effects, year fixed effects and student fixed effects (see Rockoff (2004) for a discussion). Previous research has simply dropped the grade fixed effects; however, given that this paper's focus is differences between grades, the exclusion of grade fixed effects seems ill advised. Nevertheless, I report estimates from the student-fixed effects model (excluding grade fixed effects) with the understanding that the restricted sample estimates are identified off of very few students.

Estimates based on the above 3 models (as well as the preferred specification from the text) are shown in Table 1.10. Results are extremely similar for the lagged model, the instrumented lagged model and the gains model. When including student fixed effects for the restricted sample, grade-specific experience appears to be much more important and the impact of general experience is not statistically significant. The differences in results for the student fixed effect model could be due to the exclusion of grade fixed effects or the controls for unobserved fixed student ability; however, I find it most likely that the differences are due to the lack of repetitive student observations because when estimated on the full sample of teachers, the student fixed effect model yields similar results to the other models. Table 1.11 shows estimates of the same models for the full sample of teachers. When using the entire sample, all four value-added models yield qualitatively similar results.

[21] If a student is observed for three years in the data, this student is only observed for three years in the restricted sample if he/she has a fully observed teacher in every year. Only 16% of the students who are observed for three years in the full sample, appear for three years in the restricted sample.

Table 1.8: Impact of Teacher Experience on Student Performance

	Math			Reading		
Fixed Effects:	Teacher-by-School (1)	School (2)	Teacher-by-School-by-Grade (3)	Teacher-by-School (4)	School (5)	Teacher-by-School-by-Grade (6)
Grade-specific experience	0.0185**	0.0310**	0.0261**	0.0096**	0.0172**	0.0101*
	(0.0037)	(0.0031)	(0.0050)	(0.0035)	(0.0026)	(0.0047)
(Grade-specific experience)^2	-0.0024**	-0.0036**	-0.0032**	-0.0015*	-0.0022**	-0.0016*
	(0.0007)	(0.0007)	(0.0008)	(0.0007)	(0.0006)	(0.0008)
(Grade-specific experience)^3	0.0001†	0.0002**	0.0001*	0.0001	0.0001**	0.0001
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Experience	0.0303**	0.0278**	0.0222**	0.0175**	0.0194**	0.0181**
	(0.0044)	(0.0036)	(0.0051)	(0.0041)	(0.0029)	(0.0048)
Experience^2	-0.0067**	-0.0053**	-0.0057**	-0.0043**	-0.0036**	-0.0043**
	(0.0009)	(0.0008)	(0.0009)	(0.0008)	(0.0007)	(0.0009)
Experience^3	0.0004**	0.0003**	0.0003**	0.0003**	0.0002**	0.0003**
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0000)	(0.0001)
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Grade fixed effects	Yes	Yes	No	Yes	Yes	No
Student characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Peer characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1835220	1835220	1835220	1827554	1827554	1827554

† Significant at 10%; * significant at 5%; ** significant at 1%. Standard errors clustered at class level reported in parentheses.

Notes: The dependent variable is test score. Teachers with out-of-sample experience are included and the distribution of grades taught out-of-sample is imputed as laid out in the text.

Table 1.9: Falsification: "Impact" of Future Teacher Experience on Current Student Performance

	Math			Reading		
Fixed Effects:	Teacher-by-School (1)	School (2)	Teacher-by-School-by-Grade (3)	Teacher-by-School (4)	School (5)	Teacher-by-School-by-Grade (6)
Lead teacher grade exper.	0.0006	-0.0000	0.0008	0.0025	0.0012	0.0026
	(0.0030)	(0.0033)	(0.0030)	(0.0032)	(0.0033)	(0.0033)
(Lead teacher grade exper.)^2	0.0001	0.0001	0.0001	-0.0003	-0.0001	-0.0003
	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)
(Lead teacher grade exper.)^3	-0.0000	-0.0000	-0.0000	0.0000	-0.0000	-0.0000
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Lead teacher exper.	0.0083*	0.0077*	0.0083*	0.0047	0.0043	0.0040
	(0.0035)	(0.0039)	(0.0035)	(0.0038)	(0.0038)	(0.0038)
(Lead teacher exper.)^2	-0.0021**	-0.0015†	-0.0021**	-0.0016†	-0.0010	-0.0014
	(0.0008)	(0.0009)	(0.0008)	(0.0008)	(0.0009)	(0.0009)
(Lead teacher exper.)^3	0.0001**	0.0001	0.0001**	0.0001*	0.0001	0.0001†
	(0.0000)	(0.0001)	(0.0000)	(0.0001)	(0.0001)	(0.0001)
Observations	815869	815869	815869	812774	812774	812774

† Significant at 10%; * significant at 5%; ** significant at 1%. Standard errors clustered at class level reported in parentheses.

Notes: The dependent variable is test score in period t. Lead experience and lead grade-specific experience correspond to the characteristics of student i's teacher in period t+1. All period t covariates are included in this regression.

Table 1.10: Robustness Across Value-Added Models

	Math				Reading			
Model:	Lagged IV (1)	Gains (2)	Lagged (3)	Student Fixed Effect (4)	Lagged IV (5)	Gains (6)	Lagged (7)	Student Fixed Effect (8)
Grade-specific experience	0.0211*	0.0216**	0.0212*	0.0490**	-0.0013	0.0006	-0.0009	0.0070
	(0.0088)	(0.0073)	(0.0084)	(0.0081)	(0.0084)	(0.0066)	(0.0079)	(0.0094)
(Grade-specific experience)^2	-0.0033	-0.0041*	-0.0034	-0.0091**	0.0017	0.0005	0.0015	-0.0004
	(0.0024)	(0.0020)	(0.0023)	(0.0023)	(0.0024)	(0.0018)	(0.0022)	(0.0027)
(Grade-specific experience)^3	0.0002	0.0002	0.0002	0.0006**	-0.0002	-0.0001	-0.0002	0.0001
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0002)	(0.0002)
Experience	0.0453**	0.0443**	0.0451**	0.0130	0.0624**	0.0555**	0.0613**	0.0555**
	(0.0130)	(0.0112)	(0.0124)	(0.0150)	(0.0128)	(0.0099)	(0.0120)	(0.0173)
Experience^2	-0.0079**	-0.0069**	-0.0077**	-0.0028	-0.0088**	-0.0068**	-0.0084**	-0.0050*
	(0.0021)	(0.0017)	(0.0020)	(0.0020)	(0.0020)	(0.0015)	(0.0019)	(0.0023)
Experience^3	0.0005**	0.0004**	0.0005**	0.0001	0.0006**	0.0004**	0.0006**	0.0002
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0002)
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grade fixed effects	Yes	Yes	Yes	No	Yes	Yes	Yes	No
Student characteristics	Yes	Yes	Yes	No	Yes	Yes	Yes	No
Peer characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	560874	560874	560874	560874	558451	558451	558451	558451

† Significant at 10%; * significant at 5%; ** significant at 1%. Standard errors clustered at class level reported in parentheses.

Notes: The dependent variable is test score. Headings refer to model type. The Lagged IV model is the preferred specification used throughout the main text. The gains, lagged and student fixed effect models correspond to equations 9, 10, and 11 respectively

Table 1.11: Robustness Across Value-Added Models - Full Sample

	Math				Reading			
Model:	Lagged IV (1)	Gains (2)	Lagged (3)	Student Fixed Effect (4)	Lagged IV (5)	Gains (6)	Lagged (7)	Student Fixed Effect (8)
Grade-specific experience	0.0184**	0.0182**	0.0171**	0.0290**	0.0094**	0.0090**	0.0063*	0.0173**
	(0.0037)	(0.0037)	(0.0034)	(0.0023)	(0.0035)	(0.0034)	(0.0030)	(0.0026)
(Grade-specific experience)^2	-0.0023**	-0.0022**	-0.0019**	-0.0032**	-0.0015*	-0.0014*	-0.0006	-0.0014**
	(0.0007)	(0.0007)	(0.0007)	(0.0005)	(0.0007)	(0.0007)	(0.0006)	(0.0005)
(Grade-specific experience)^3	0.0001†	0.0001	0.0000	0.0001**	0.0001	0.0001	0.0000	0.0001†
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Experience	0.0302**	0.0309**	0.0352**	0.0324**	0.0176**	0.0182**	0.0229**	0.0215**
	(0.0044)	(0.0043)	(0.0040)	(0.0029)	(0.0041)	(0.0040)	(0.0035)	(0.0033)
Experience^2	-0.0067**	-0.0067**	-0.0069**	-0.0067**	-0.0043**	-0.0043**	-0.0045**	-0.0045**
	(0.0009)	(0.0009)	(0.0008)	(0.0006)	(0.0008)	(0.0008)	(0.0007)	(0.0007)
Experience^3	0.0004**	0.0004**	0.0004**	0.0004**	0.0003**	0.0003**	0.0002**	0.0002**
	(0.0001)	(0.0001)	(0.0001)	(0.0000)	(0.0001)	(0.0000)	(0.0000)	(0.0000)
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grade fixed effects	Yes	Yes	Yes	No	Yes	Yes	Yes	No
Student characteristics	Yes	Yes	Yes	No	Yes	Yes	Yes	No
Peer characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1834877	1834877	1834877	1834877	1827210	1827210	1827210	1827210

† Significant at 10%; * significant at 5%; ** significant at 1%. Standard errors clustered at class level reported in parentheses.

Notes: The dependent variable is test score. Headings refer to model type. The Lagged IV model is the preferred specification used throughout the main text. The gains, lagged and student fixed effect models correspond to equations 9, 10, and 11 respectively

Table 1.12: Full Set of Coefficients from Table 1.4

	Math			Reading		
Fixed Effects:	School (1)	Teacher-by-School (2)	Teacher-by-School-by-Grade (3)	School (4)	Teacher-by-School (5)	Teacher-by-School-by-Grade (6)
Female student	-0.0014	-0.0018	-0.0020	-0.0041	-0.0048	-0.0043
	(0.0031)	(0.0031)	(0.0031)	(0.0033)	(0.0033)	(0.0033)
Black student	-0.0206**	-0.0200**	-0.0197**	-0.0385**	-0.0438**	-0.0377**
	(0.0032)	(0.0032)	(0.0032)	(0.0033)	(0.0032)	(0.0033)
Hispanic student	0.0229**	0.0186*	0.0239**	-0.0036	-0.0056	-0.0021
	(0.0076)	(0.0077)	(0.0076)	(0.0076)	(0.0076)	(0.0076)
Parental education						
High school dropout	-0.0231**	-0.0138**	-0.0243**	-0.0402**	-0.0334**	-0.0412**
	(0.0043)	(0.0044)	(0.0043)	(0.0047)	(0.0048)	(0.0047)
College graduate	0.0183**	0.0090**	0.0193**	0.0101**	0.0017	0.0111**
	(0.0032)	(0.0032)	(0.0032)	(0.0033)	(0.0033)	(0.0033)
Student on subsidized lunch	-0.0290**	-0.0298**	-0.0286**	-0.0259**	-0.0261**	-0.0259**
	(0.0026)	(0.0026)	(0.0026)	(0.0028)	(0.0028)	(0.0028)
Limited english proficiency	0.1207**	0.1229**	0.1200**	0.1516**	0.1534**	0.1528**
	(0.0084)	(0.0085)	(0.0084)	(0.0092)	(0.0093)	(0.0092)
Student-teacher same sex	0.0050†	0.0055†	0.0053†	-0.0066*	-0.0071*	-0.0066*
	(0.0030)	(0.0031)	(0.0030)	(0.0032)	(0.0033)	(0.0032)
Student-teacher same ethnicity	0.0022	-0.0046	0.0009	-0.0101	0.0113†	-0.0109
	(0.0086)	(0.0071)	(0.0086)	(0.0091)	(0.0064)	(0.0091)
Class size	-0.0051**	-0.0032**	-0.0050**	-0.0036**	-0.0018**	-0.0037**
	(0.0006)	(0.0006)	(0.0007)	(0.0006)	(0.0005)	(0.0007)
Teacher was switched this year	0.0142†	0.0209**	0.0182*	-0.0009	0.0005	0.0033
	(0.0079)	(0.0078)	(0.0084)	(0.0076)	(0.0066)	(0.0081)

Continued on next page

Table 1.12: Continued from previous page

	Math			Reading		
Fixed Effects:	Teacher-by-School	School	Teacher-by-School-by-Grade	Teacher-by-School	School	Teacher-by-School-by-Grade
Total number of times switched	0.0007	-0.0184**	-0.0029	-0.0039	-0.0095**	-0.0010
	(0.0063)	(0.0031)	(0.0105)	(0.0062)	(0.0026)	(0.0100)
Experience	0.0453**	0.0347**	0.0436*	0.0624**	0.0321**	0.0647**
	(0.0129)	(0.0076)	(0.0190)	(0.0126)	(0.0064)	(0.0183)
Experience^2	-0.0079**	-0.0055**	-0.0076**	-0.0088**	-0.0058**	-0.0094**
	(0.0021)	(0.0020)	(0.0023)	(0.0020)	(0.0017)	(0.0022)
Experience^3	0.0005**	0.0004*	0.0005**	0.0006**	0.0004**	0.0007**
	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(0.0002)
Grade-specific experience	0.0211*	0.0374**	0.0325†	-0.0013	0.0083	0.0075
	(0.0087)	(0.0082)	(0.0171)	(0.0083)	(0.0070)	(0.0166)
(Grade-specific experience)^2	-0.0033	-0.0071**	-0.0039	0.0017	-0.0010	0.0022
	(0.0024)	(0.0023)	(0.0026)	(0.0023)	(0.0020)	(0.0025)
(Grade-specific experience)^3	0.0002	0.0004†	0.0001	-0.0002	0.0000	-0.0002
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Second School	-0.0152	0.0151	-0.0282	0.0703	0.0197	0.0848†
	(0.0464)	(0.0245)	(0.0472)	(0.0472)	(0.0213)	(0.0514)
2nd Schl X Exp.	0.0129	0.0015	-0.0270	-0.0233	-0.0122	-0.0477
	(0.0230)	(0.0205)	(0.0317)	(0.0228)	(0.0179)	(0.0316)
2nd Schl X Exp.^{2}	-0.0013	0.0008	-0.0006	0.0078	0.0042	0.0100†
	(0.0050)	(0.0046)	(0.0054)	(0.0050)	(0.0041)	(0.0055)
2nd Schl X Exp.^{3}	0.0001	-0.0000	-0.0000	-0.0006†	-0.0003	-0.0007†
	(0.0003)	(0.0003)	(0.0004)	(0.0003)	(0.0003)	(0.0004)
2nd Schl X Grd. Exp.	-0.0221	-0.0251*	0.0205	-0.0288*	-0.0250*	0.0072
	(0.0136)	(0.0122)	(0.0237)	(0.0130)	(0.0105)	(0.0236)
2nd Schl X (Grd. Exp.)^2	0.0042	0.0054	0.0033	0.0029	0.0047	-0.0022
	(0.0039)	(0.0037)	(0.0041)	(0.0038)	(0.0032)	(0.0041)

Continued on next page

Table 1.12: Continued from previous page

	Math			Reading		
Fixed Effects:	Teacher-by-School	School	Teacher-by-School-by-Grade	Teacher-by-School	School	Teacher-by-School-by-Grade
2nd Schl X (Grd. Exp.)^2	-0.0002	-0.0004	-0.0002	0.0001	-0.0002	0.0004
	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
Third School	0.3533*	0.2727	0.3720†	-0.0071	0.1788	0.0172
	(0.1697)	(0.1749)	(0.1946)	(0.1813)	(0.1486)	(0.2185)
3rd Schl X Exp.	-0.1880†	-0.1005	-0.1814	-0.0449	-0.0760	-0.0549
	(0.1023)	(0.1033)	(0.1184)	(0.1073)	(0.0897)	(0.1277)
3rd Schl X Exp.^{2}	0.0274	0.0140	0.0272	0.0077	0.0140	0.0054
	(0.0188)	(0.0186)	(0.0208)	(0.0195)	(0.0163)	(0.0222)
3rd Schl X Exp.^{3}	-0.0012	-0.0007	-0.0012	-0.0003	-0.0009	-0.0003
	(0.0011)	(0.0010)	(0.0012)	(0.0011)	(0.0009)	(0.0012)
3rd Schl X Grd. Exp.	0.0113	-0.0527	-0.0042	-0.0019	-0.0289	0.0076
	(0.0361)	(0.0321)	(0.0512)	(0.0362)	(0.0277)	(0.0511)
3rd Schl X (Grd. Exp.)^2	0.0033	0.0150	0.0068	0.0053	0.0070	0.0105
	(0.0107)	(0.0096)	(0.0119)	(0.0106)	(0.0083)	(0.0119)
3rd Schl X (Grd. Exp.)^3	-0.0006	-0.0011	-0.0008	-0.0008	-0.0005	-0.0011
	(0.0008)	(0.0008)	(0.0009)	(0.0008)	(0.0007)	(0.0009)
Peer Characteristics						
Ave. Lagged math score	-0.0313*	-0.2573**	0.0240†			
	(0.0122)	(0.0080)	(0.0137)			
Ave. Lagged read score				0.1059**	-0.1570**	0.1673**
				(0.0114)	(0.0067)	(0.0126)
Fraction female	0.0421*	0.0311†	0.0332†	-0.0158	0.0212	-0.0273
	(0.0181)	(0.0178)	(0.0189)	(0.0181)	(0.0154)	(0.0189)

Continued on next page

Table 1.12: Continued from previous page

	Math			Reading		
Fixed Effects:	Teacher-by-School	School	Teacher-by-School-by-Grade	Teacher-by-School	School	Teacher-by-School-by-Grade
Fraction black	-0.0319	-0.2291**	0.0041	0.0938**	-0.1076**	0.1298**
	(0.0219)	(0.0200)	(0.0231)	(0.0216)	(0.0171)	(0.0228)
Fraction hispanic	-0.0704†	-0.2057**	-0.0477	-0.1023**	-0.1967**	-0.0667†
	(0.0376)	(0.0320)	(0.0397)	(0.0376)	(0.0296)	(0.0396)
Frac. parent ed: Dropout	-0.0150	0.0492**	-0.0175	0.0261	0.0426**	0.0274
	(0.0197)	(0.0170)	(0.0205)	(0.0197)	(0.0149)	(0.0206)
Frac. parent ed: College grad.	0.0113	0.0096	0.0051	-0.0113	-0.0069	-0.0192
	(0.0145)	(0.0123)	(0.0152)	(0.0140)	(0.0110)	(0.0148)
Frac. subsidized lunch	0.0081	-0.0833**	0.0255*	0.0411**	-0.0503**	0.0510**
	(0.0116)	(0.0109)	(0.0123)	(0.0115)	(0.0094)	(0.0122)
Frac. limited english proficient	0.1572**	0.0719†	0.1507**	0.2427**	0.0678†	0.2847**
	(0.0433)	(0.0391)	(0.0454)	(0.0439)	(0.0363)	(0.0464)
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Grade fixed effects	Yes	Yes	No	Yes	Yes	No
Observations	560874	560874	560874	558451	558451	558451

† Significant at 10%; * significant at 5%; ** significant at 1%. Clustered standard errors reported in parentheses

Notes: The dependent variable is test score. This table provides the full set of coefficients for the regressions presented in Table 1.4. The exact coefficients for many variables in these regressions should not be interpreted causally because as in Rockoff (2004), I do not attempt to credibly identify the impact of exogenous changes in these controls and am only including these variables to control for potentially confounding factors.

Chapter 2: The Role of Peers and Grades in Determining Major Persistence in the Sciences

2.1 Introduction

Lagging persistence in the sciences has been a major concern for policy makers over the past forty years. In particular, much of the research and policy on this topic has focused on improving the representation of females and minorities in science majors. While frequently lumped together into the "science" category, students in the physical and life sciences exhibit dramatically different patterns of major intent and persistence. This paper highlights the importance of disaggregating the sciences into physical and life sciences, both in conducting research and implementing policy.

One difference between this study and much of the previous literature is the selectivity of the school analyzed and the fact that it is a major research university. The findings of nationally representative studies may not reflect trends affecting top universities and these schools have a disproportionate impact. Much of the reason that policy makers are deeply concerned about science graduation rates is that it is perceived to impair global competitiveness and reduce the potential benefits of positive externalities associated with scientific research. In this regard, a selective research university is a particularly important setting to understand trends and determinants of major persistence because these students are disproportionately likely to contribute to scientific progress.

Using administrative data from a large selective research university (LSRU herein), I examine the determinants of entering and then persisting in physical and life science majors. Because physical and life science persistence patterns are not yet well understood at selective universities, I devote considerable time to descriptive analysis, particularly in examining gender and ethnicity differentials. I show, for example, that failing to account for differences between physical and life sciences may lead to spurious findings regarding the importance of gender because females are over represented in the life sciences, where attrition is particularly high.

In addition to describing persistence trends, this paper investigates the impact of one's peers on

major persistence. Peer quality has been shown to be an important determinant of student performance in a variety of settings (Carrell, Fullerton and West, 2009; Sacerdote, 2001; Zimmerman, 2003). However, there is limited information regarding the influence of peers on major persistence and choice. Recent studies using randomly assigned roommates for identification such as Sacerdote (2001) and Han and Li (2009) find no evidence of residential peer influence on major choice. Instead of examining residential peer effects, I investigate the existence of peer effects in one's courses. For the physical sciences, I find evidence of positive peer effects in one's core physical science classes suggesting that classmates may have a larger influence on academic decisions than roommates.

The importance of grades in determining course choice has been documented extensively (Bar, Kadiyali and Zussman, 2009; Fournier and Sass, 2000; Sabot and Wakeman-Linn, 1991). Given that grade inflation has disproportionately affected non-science fields, a grading gap has emerged that provides students with an incentive to defect from the sciences. Furthermore, there is evidence that females respond more strongly to grade incentives than do males, potentially exacerbating the persistence gap between men and women (Owen, 2010; Rask and Tiefenthaler, 2008).

While these phenomena affect all sciences, there exists a large amount of variation in grading standards within the sciences as well. In fact, for my data, the gap between average grades in the life and physical sciences is nearly as large as the gap between life and non-sciences. In both fields, I find that students are less likely to persist as their non-science grades improve and more likely to persist as their own field grades improve. Furthermore, I confirm that females appear more sensitive to grades; however, this differential sensitivity is limited to the physical sciences, where females are a minority group.

Since many of the factors affecting persistence are similar in the physical and life sciences, research analyzing all science majors together will in many cases arrive at qualitatively correct conclusions. However, the large body of research on the gender persistence gap need be particularly careful to distinguish between physical and life scientists.

2.2 Data

For this study, I use longitudinal administrative data from a large selective research university. This data encompasses the entire universe of this university, including the complete transcript of courses and grades for every entering student from 1997-2003.[1] These transcripts also include unique course identifiers; thus, I can exactly identify each student's peers in every course she takes. Course identifiers are constant over time facilitating the inclusion of course fixed effects. The final cohort (2003) is followed until 2008 when most (but not all) students have either graduated or withdrawn. This transcript data is matched to admissions data such as SAT scores, class rank and demographic information.[2]

Students entering this university are not representative of college students nationally. As shown in column 1 of Table 2.1, 53.7% of students in this period are white, 4.6% are black, 16.1% are Asian, 5.4% are Hispanic, 2.2% report two races and 17.2% of students fail to report a race or ethnicity. Slightly over 50% of entering students are male. The average SAT score over this period is 1358.2 and more than 75% of incoming students were in the top 10% of their graduating high school class. 33% of students enter LSRU with credit for at least one college course and 24% enter with credit for calculus.[3] While this university is clearly not nationally representative, it is fairly well representative of students who are likely to eventually perform top level research in the sciences.

As part of the admissions process at LSRU, students are asked to indicate their intended major. While students in the liberal arts sector of LSRU can list either science or non-science majors, the majority of students at LSRU are admitted to a specific branch and must major within that branch (e.g. students admitted into the college of engineering must intend to major in STEM). Of the 17,145 students who enter LSRU during this time period, 1,634 are either missing an intended major or list "undeclared." For all analyses that require an intended major, these observations are

[1] One exception is transfer students who are generally not included in this data.

[2] Unfortunately, course instructors have not been matched to student transcripts for this university and historical faculty information is limited. A lengthy attempt to match professors to classrooms only successfully matched 30% of courses.

[3] In most cases, college credit is obtained by taking AP/IB courses in high school.

dropped.[4] Of students who list an intended major, 39% intend to major in a non-STEM field, 37% intend to major in the physical sciences, and 24% intend to major in the life sciences.

Columns 2-5 of Table 2.1 break out descriptive statistics by intended major. Students who intend to major in the physical or life sciences are generally stronger students in terms of SAT scores, incoming college credits and high school rank. The physical sciences are over 70% male whereas the life sciences are over 60% female. Asian students are overrepresented in the physical sciences whereas black and Hispanic students are under represented. Based on these minimal descriptive statistics, it is already clear that the life and physical sciences attract different types of students.

2.3 Empirical Methods and Results

Graduating with a science degree is the result of first intending to major in a science field, and second, persisting in this field.[5] I therefore first explore patterns of intended major and then, conditional on intended major, I examine patterns of persistence.

2.3.1 Intended Major Choice

Since different races and genders may have different preparation levels on average, I use a regression framework to better understand the type of student who intends to major in life or physical sciences. Importantly, I am not using regression in an attempt to identify causal estimates, but rather to refine the descriptive analysis. My primary question is whether the gaps between genders and races documented above can be explained by differential preparation. To address this question I use the multinomial logit function shown by equation 1 to estimate which factors

[4]Including these students and using an imputed major has little impact on results.

[5]While it is possible to switch to a science major after intending a non-science major, in practice less than 5% of non-science majors transfer to become science majors at LSRU.

contribute to declaring a life science, physical science or a non-science major.

$$P(y = j) = \frac{e^{\beta' \theta_j}}{1 + e^{\beta' \theta_j}} \qquad for\ j = 0, 1, 2 \tag{1}$$

Table 2.2 shows results from the multinomial logit regression where the choices are intending to major in physical science, life science, or non-science. Consistent with the findings of Turner and Bowen (1999), many of the gender and ethnicity gaps remain when controlling for various measures of high school preparation. Controlling for observables, females are 24.2 percentage points less likely to intend to major in physical sciences and 13.4 percentage points more likely to major in life sciences. Asian students are 9.4 percentage points more likely to major in physical sciences and 2.4 percentage points less likely to major in life sciences. The magnitudes of these gaps is considerably smaller than the raw gaps, suggesting that high school preparation (or ability) accounts for some, but not all, of the intended major choices of females and Asians.

For Hispanic and black students, however, the raw gap in the likelihood of declaring a physical or life science major disappears after controlling for preparation/ability, suggesting that differences in intended majors between black/Hispanic students and white students is entirely due to differential preparation between these groups. If anything, conditional on high school performance and preparation, black and Hispanic students are more likely than white students to pursue physical or life science majors.

Consistent with the notion that stronger students tend to enter physical sciences, a one standard deviation increase in SAT score is associated with an 8.3 percentage point increase in the likelihood of intending to major in the physical sciences and a 2.8 percentage point decrease in the likelihood of intending to major in the life sciences. One's percentile in high school, on the other hand, has little effect on the propensity to major in physical science and increases the chance of intending to major in life science by 3.2 percentage points.

As expected, entering college with calculus credit strongly influences the probability of intending a physical science major, improving the likelihood by 16.6 percentage points. This effect may

be because students with calculus credit have a preference for math intensive fields or because students with calculus credit have (or feel like they have) solid preparation to pursue math intensive curricula. Controlling for whether a student has taken AP/IB calculus, the raw number of incoming college credits a student has is associated with a small decrease in the likelihood of majoring in the physical sciences. This makes intuitive sense because a student with only calculus credit is pulled strongly towards math intensive fields whereas a student with calculus credit, English credit, and biology credit may have more attractive options outside of the physical sciences. The impact of incoming calculus credit and total credits is reversed for the life sciences; however, the magnitudes of the effects are much smaller than for physical science.

Based on the determinants of entering each field, physical and life sciences are less similar than are life and non-science. In cases where data limitations or sample sizes prevent separately analyzing life, physical and non-scientists, the above regression provides some evidence that the appropriate grouping may be physical science vs life and non-science instead of non-science vs physical and life science.

2.3.2 Descriptive Analysis of Persistence

Conditional on intending to major in physical or life science, there is substantial variation in the probability of graduating with a degree in life or physical sciences across groups. Because the vast majority of students at LSRU eventually earn a degree, students who fail to persist in physical or life sciences are generally switching to an alternative major. As with the initial decision to declare a physical or life science major, the persistence patterns in the life and physical sciences are very different.

2.3.2.1 Gender Gap

The most dramatic difference between life and physical science persistence patterns is that at LSRU university, the gender gap in persistence is *solely* driven by a gender gap in the physical

sciences. This fact is demonstrated emphatically by Figures 2.1 and 2.2.[6] As can be seen in Figure 2.1, there is essentially no difference in the persistence rates of females vs males in the life sciences. Not only do the two genders have similar graduation rates, but the trajectory with which the two groups attrit is virtually identical. Conversely, the figure for physical science shows a substantial gap between the persistence rates for males and females. Over 80% of males who intend to major in physical science successfully due so, whereas only 70% of females who intend to major in physical science persist. If one were to combine physical and life sciences in analysis, a persistence gap would appear, both because a persistence gap exists in the physical sciences and because raw persistence rates are much lower in the life sciences where females are overrepresented.

2.3.2.2 Race and Ethnicity

The persistence gaps between black students and non-black students are qualitatively similar for life and physical sciences. Figures 2.3 and 2.4 show that black students are nearly 20 percentage points less likely to persist in both the physical and life sciences. The difference in persistence between Hispanic students and non-Hispanic students is smaller but qualitatively similar to the black/non-black gap and is shown in Figures 2.5 and 2.6. Figures 2.7 and 2.8 show persistence rates broken out by whether a student is Asian or not. Asian students are considerably more likely to persist in the physical sciences; however, their persistence pattern is similar to non-Asian students for the life sciences.

2.3.2.3 SAT Scores

Students with higher SAT scores are more likely to persist in the sciences. This is true for both physical and life science and the relationship between SAT scores and persistence is similar across

[6] I examine 4 year persistence rather than 6 year persistence in order to be able to include the 2002 and 2003 cohorts in all analyses. The qualitative results are identical when using 6 year persistence rates.

fields. Figures 2.9 and 2.10 show persistence rates separately for three SAT categories at LSRU. The "high SAT" category is the top quartile of SAT scores, the "low SAT" category is the bottom quartile and the "mid SAT" category represents students with SAT scores in the interquartile range.

2.3.3 Empirical Methods: Determinants of Persistence

While the above descriptive analysis highlights the raw persistence gaps between various groups, it fails to control for factors which may be correlated with group membership. As in equation 1, I control for observable factors which proxy for high school preparation. Furthermore, I can control for both observed and unobserved characteristics of students once they arrive. My aim in these analyses is to investigate the role of preparation, peers and grades.

Since the vast majority of major changes are from life or physical science majors to non-science fields, I focus on the probability of persisting in one's field rather than attempting to separately investigate transitions within and out of science majors. When a linear probability model is estimated, a small fraction of students have predicted persistence rates of above 100% and thus I opt to estimate a logit model instead. Specifically, I estimate

$$P(y_{ij} = 1) = \frac{e^{\beta' X}}{1 + e^{\beta' X}} \tag{2}$$

where $\beta' X = \beta_1 GPAs_{ij} + \beta_2 Peers_{ij} + \beta_3 X_i + \gamma_j$

y_{ij} is an indicator for whether student i, taking course j, persists in her major field through the fourth year. The vector $GPAs$ includes each student's overall GPA as well as separate GPAs calculated for physical science courses and life science courses. The $Peers$ vector includes a measure of average peer quality for student i in course j.[7] A detailed explanation of the peer quality measure is given in Section 2.3.4.3. The vector X_i includes a variety of fixed student characteristics, specifically SAT score, high school percentile, race, sex, incoming college credits

[7] Student i's peer characteristics are calculated excluding student i.

and an indicator for having calculus credit prior to entering college. Naturally, within individual i, cross course correlations will not be zero, thus all standard errors are clustered at the student level.

This model is estimated separately for life and physical sciences. For regressions predicting persistence in physical sciences, I include only core required physical science courses in the regression. Similarly, regressions that predict persistence in the life sciences only include core required life science courses. Focusing on large core courses facilitates the inclusion of course fixed effects and limits the analysis only to students who plausibly intend to major in their listed "intended major."[8]

Including course fixed effects has two major advantages. First, students in the same course presumably have similar interests and thus the inclusion of course fixed effects controls for unobserved differences between students who choose different courses. Second, because the same courses are offered each year, I am able to identify the effect of peer characteristics using across time variation, holding course topic fixed. Without course fixed effects, the impact of peers would be confounded by the fact that one's peers would be determined by one's course choices (Sacerdote, 2001). Cohort fixed effects are included in some specifications in order to account for time varying unobserved factors that affect an entire cohort's persistence.

2.3.4 Results: Determinants of Persistence

2.3.4.1 Physical Science

Table 2.3 shows the coefficients when equation 2 is estimated for physical science majors. The first column omits the course fixed effects and the last column introduces cohort fixed effects. For ease of interpretation the average marginal effect is given. When controlling for performance factors, females still are more likely to drop out of the physical sciences but the magnitude is much smaller than that implied by the raw gap shown in Figure 2.2. The raw persistence gap is nearly 10 percentage points and this gap drops to just 2.7 percentage points when controlling for performance

[8] A small percentage of students list a science major as their intended major during admissions but take no science courses during their first year. These students are omitted from the persistence analysis.

and course fixed effects. The inclusion of course fixed effects may understate the persistence gap if being a female causes course choices which lead to lower persistence. However, column (1) of Table 2.3 shows that even when omitting the course fixed effect the gender gap in persistence is just 2.8 percentage points.

The persistence gap between black and white students is statistically and substantively insignificant once controlling for other factors. This is rather remarkable given that the raw gap in persistence is 12.4 percentage points. This phenomenon is true for Hispanic students as well, as the large raw gap is completely explained by other factors. Importantly, some of the other factors for which I control, such as GPA, may themselves be a function of one's gender or race and thus the regression might be over controlling. Regardless, the results indicate that once other factors are equalized, being black or Hispanic has no impact on persistence. Asian students are more likely to persist, controlling for other factors. When controlling for course and cohort fixed effects, Asian students are 3.7 percentage points more likely to persist than comparable white students.

High school preparation generally is a weak predictor of persistence once college grades are controlled for. This does not mean that preparation is unimportant, rather that the entire benefit to high school preparation is captured in performance in college courses. Taking calculus before entering college is associated with a 2.8 percentage point increase in persistence probabilities. However, the positive effect of calculus is only statistically significant when controlling for course and cohort fixed effects and is not robust across specifications. Given that my preferred specification is to include course and cohort fixed effects, I view this result as providing some suggestive evidence that calculus preparation may be helpful for persistence.

Performance in college courses is a critical determinant of persistence. As shown in column (3) of Table 2.3, a one point increase in a student's average physical science grades is associated with an 11.5 percentage point increase in the probability of persisting in the physical sciences.[9] Controlling for a student's grade performance in her science courses, improving overall GPA by one point decreases persistence by 3.5 percentage points. There may be small negative effects to

[9]A one point increase is the equivalent of a one letter grade increase e.g. improving from a B to an A.

performing well in life science courses, but this is not statistically significant. While the above results may partially reflect students simply gravitating towards their relative strengths, I show in section 2.4.1 that grading standard differentials across fields may lead to student attrition that is unrelated to the revelation of relative ability.

Broadly, the inclusion of cohort and course fixed effects has little impact on the coefficients in Table 2.3. The notable exception is that the first column of Table 2.3 suggests that having a larger class is beneficial to completing the major. However columns 2-3 show that this effect is completely absent when controlling for course fixed effects. The correlation between class size and major persistence is unlikely to be causal since there is no evidence of this correlation within a course.

A measure of peer quality is controlled for throughout Table 2.3 and I explore these results in Section 2.3.4.3.

2.3.4.2 Life Science

Table 2.4 shows the coefficients when equation 2 is estimated for life science majors. Given that no raw persistence gap exists between genders for the life sciences, it is no surprise that gender is found to have no predictive power for persistence probabilities. Unlike in the physical sciences, the raw persistence gaps for Hispanic students is not fully explained by other factors. Controlling for factors such as performance and course choice, Hispanic students are still 8.2 percentage points less likely than white students to persist in the life sciences. The raw gap in persistence between Hispanic and white students in the life sciences is 20.7 percent so nearly half of this persistence gap cannot be explained by academic performance. Although the raw persistence rate for Asian students in the life sciences is only 5.85 percentage points worse than white students, this raw gap cannot be explained by performance or preparation factors.

Just as in physical sciences, high school preparation generally is a weak predictor of persistence once college grades are controlled for. Higher SAT scores are associated with slightly lower persistence rates (1.9 percentage points per standard deviation) but calculus and other entering college

credits are statistically and substantively insignificant. Given that college grades are controlled for in this regression, it is not surprising that SAT score has a slightly negative effect. One potential explanation is that students with high SAT scores but average first year college grades may lack certain study or organizational skills that make persistence difficult in life science courses which typically value these skills.[10]

The effect of own subject grades for life scientists is extremely similar to the effect for physical scientists. A one point increase in a student's GPA in her first year life science courses is associated with a 10.7 percentage point increase in the probability of persisting. Similarly higher grades in non-science courses are associated with lower persistence rates. A one point increase in a student's first year GPA holding science GPA constant is associated with an 8.1 percentage point decrease in persistence. The effect of non-science GPA on persistence in the life sciences is more than twice as large as for the physical sciences.

Somewhat surprisingly, receiving higher grades in one's physical science courses is associated with higher persistence in the life sciences. This effect partially reflects that students who perform well in their physical science courses are more skilled and thus more capable to persist in any field they choose. In addition, the pulling effect of high grades in the physical sciences is likely fairly small given that it is very rare to transfer into the physical sciences.[11] Furthermore, receiving higher grades in certain physical science courses may directly promote life science persistence because courses such as organic chemistry are technically physical science courses, but are required for many life science majors.

As with the physical sciences, there is a correlation between class size and persistence, but this disappears when controlling for course fixed effects. Students who take larger courses tend to have lower persistence rates, but this should not be interpreted causally.

[10] Anecdotal evidence suggests that study and organizational skills may be more important in the life sciences compared to physical sciences, because the former rewards factual knowledge (which requires studying) more than the latter.

[11] Fully 95 percent of successful physical science majors entered college intending to major in the physical sciences.

2.3.4.3 Peer Effects

In order to examine the effect of peers on an individual's persistence, I include a measure of peer persistence in the above regression. Were I to simply include the average persistence of one's classmates, I would overstate the importance of the peer effect because of the reflection problem (Manski, 1993). Since student i may impact her peer's persistence through her own persistence, the persistence of her peers is not exogenous to her own persistence. Previous research avoids conflating endogenous and exogenous peer effects by focusing on peer characteristics determined prior to college enrollment such as SAT score (Sacerdote, 2001). I follow a similar tact, but rather than examining a particular factor, I aggregate these factors to measure the propensity to persist.[12]

Specifically, I use the predicted values from a regression of persistence on pre-college characteristics to generate a propensity score for each student. This propensity score is a linear combination of pre-college characteristics and thus student i's propensity score will be exogenous to student j's persistence probabilities. Using the propensity score for each student, I calculate average propensity scores by class. Since these averages vary by semester, they are still identified with the inclusion of course and cohort fixed effects.

Although it seems unlikely that student i would influence student j's propensity score, it is possible that two propensity scores are jointly determined by the admissions committee. Assuming that the admissions committee has more information regarding a student's propensity to persist than I have included in my regression, any broad goal to improve persistence in a cohort may bias estimates upwards. Because of this plausible threat to identification, my preferred estimates include cohort fixed effects to control for global admissions changes that directly affect both an individual and the pre-college characteristics of her peers. In practice, omitting the cohort fixed effects yields very similar results. When estimated without course fixed effects, the magnitude on the peer coefficient increases dramatically, but this simply reflects that students sort into classrooms and thus I do not consider it as evidence of peer effects.

[12]Carrell et al. (2009) uses a similar strategy and uses pre-college characteristics to generate predicted GPA's for one's peers.

Table 2.5 shows the results for peer effects. Row 1 shows that students are more likely to persist when their peers are more likely to persist. A 10 percentage point increase in the propensity of one's peers to persist leads to a 2.08 percentage point increase in the probability of persistence. This effect decreases very slightly when cohort fixed effects are included, suggesting that joint determination of persistence within a cohort is not a major concern. The reduced form impact of peer composition in first year courses may overstate the importance of peers in a single course. Since a student is likely to have repeated interactions with first year peers, the effect I capture is the cumulative effect of these interactions.

While the existence of peer effects is interesting in its own right, there are only clear policy implications if non-linearities exist across who benefits from high quality peers. In a world with homogenous peer effects, all redistributions or reorganizations will yield an equivalent amount of spillover benefits. To investigate the possibility of non-linear peer effects, I re-estimate equation 2 on two subsamples determined by one's own propensity score. Table 2.5 shows that students who are at most risk of failing to persist are also most influenced by their peers. For the bottom quartile, a 10 percentage point increase in the propensity scores of one's peers leads to a 3.53 percentage point increase in own persistence. The effect of peers on the upper quartile is much smaller and not statistically significant. This result is consistent with the findings of Carrell et al. (2009) that low achieving students benefit most from exposure to high achieving peers.

Consistent with Han and Li (2009), I find evidence that females are more influenced by their peers than males. While both males and females benefit from exposure to higher quality peers, the effect is more than twice as large for women as compared to men. A 10 percentage point increase in the propensity scores of one's peers increases the likelihood of a female persisting by 3.70 percentage points compared to only 1.37 percentage points for males.

Unlike the physical science analysis, I find little evidence that peer effects are important in the life sciences. This may be because the substantially smaller sample size for the life sciences prevents the detection of such effects or may simply reflect a lack of peer effects in the life sciences for my sample. Anecdotal evidence and discussions with life and physical science professors

suggest that physical science courses rely more heavily on group work and collaborative problem sets, which may explain why peers appear to be more important in the physical sciences.

2.4 Discussion and Extensions

2.4.1 Grades

In both life and physical sciences, improvements in own subject GPA are associated with greater persistence in that subject whereas improvements in non-science GPA are associated with transferring away from the intended major. This intuitive finding could in fact reflect optimal sorting if grades are indicative of relative strengths. In terms of maximizing each student's potential, it makes sense for a student who initially declares a physical science major to drop out if she discovers that she is not well suited to it. In particular, if students enter college with incorrect information regarding their relative strengths, grading provides a potentially effective mechanism for informing a student which major field they should choose (Stinebrickner and Stinebrickner, 2009). Unfortunately, this mechanism is only effective when grading standards are consistent across disciplines. Table 2.6 shows that this is emphatically not the case. As is true nationally, grading standards are dramatically different between majors at LSRU.

Furthermore, the difference in average grades is not simply a reflection of differential student sorting into various majors. As shown in the third column of Table 2.6, students who intend to major in physical sciences receive higher grades in their non-science courses compared to their physical science courses. One might expect that physical science majors would perform best in their physical science courses, but the opposite is true. As a result, a student who is best suited for life or physical sciences may in fact choose a non-science major if she is attracted by her higher average grades in these fields.

2.4.1.1 Differential Gender Responses to Grades

Rask and Tiefenthaler (2008) provides evidence that females are more sensitive to grades in

determining economics major persistence and proposes that a similar phenomenon may contribute to the gender gap in the sciences. To investigate this possibility I re-estimate the above model separately for females and males. For the physical sciences, where a gender gap exists, the correlation between grades and persistence is very consistent with the hypothesis put out by Rask and Tiefenthaler.

As shown in Table 2.7, in the physical sciences, females are more sensitive to grades both in terms of major field performance and outside option performance. For females a one point increase in GPA in physical science courses improves the probability of persistence by 13.4 percent whereas the corresponding figure for males is only 10.7 percent. Similarly, a one point increase in overall GPA (holding constant science GPA) leads to a much steeper decline in persistence for females than for males. Interestingly, the same does not appear to be true for the life sciences. Own field grades have a similar impact on both males and females and outside options have a larger impact for males than they do for females.

While the reason for a grade response differential in the physical sciences but not the life sciences is unclear, one possibility is that student grade response is a function of perceived "minority" status as opposed to gender per se. There is insufficient evidence presented in this study to conclude anything about the causes of gender response differentials, but this study's results as well as the findings of Rask and Tiefenthaler (2008) are consistent with the social psychology theories of stereotype vulnerability or attributional ambiguity (Crocker and Major, 1989). Essentially, these theories hold that individuals in a minority position have a tendency to be influenced by stereotypes about one's social category. In particular, Aronson and Inzlicht (2004) find that students who exhibit signs of stereotype vulnerability are less able to gauge performance and consequently have unstable academic self-concept and efficacy. Furthermore, the authors note that "unstable efficacy is associated with increased sensitivity to performance feedback, both positive and negative" (p. 834). In other words, a female majoring in the physical sciences may have a particularly large response to grades because she is in the minority whereas females majoring in the life sciences are not a minority group.

2.4.2 Peers

While recent influential research on peer effects finds no impact on major choices this may be due to looking at the wrong peers (Foster, 2006; Stinebrickner and Stinebrickner, 2006). Previous research obtains convincing identification through the random assignment of roommates, but it is far from clear that roommates are the peers who would be expected to impact major persistence. While my identification strategy is not as perfectly clean, my study benefits in that it examines peer effects in an important academic context. Naturally, the impact of one's roommates may have a different impact on major persistence than the impact of one's classmates, particularly one's classmates in key major courses.

Because the benefit of persistent peers is non-linear, social gains are possible by sorting students efficiently. Since students with a low propensity to persist benefit from exposure to high propensity peers and high propensity students do not seem to be brought down by low propensity peers, complete integration yields optimal results. While theoretically the implication of the non-linear peer effect is clear, in practice the impact of resorting may yield unanticipated consequences not captured in my analysis. The potential for these unanticipated consequences is highlighted by Carrell, Sacerdote and West (Unpublished). Although Carrell et al. (2009) find evidence of non-linear peer effects similar to those found in this paper, Carrell, Sacerdote and West (Unpublished) documents that an attempt to exploit these non-linearities to improve the outcomes of low achievers actually negatively impacted these students.

The finding that females are more susceptible to peer influence than males is consistent with a large body of literature in social psychology (Eagly, 1978) in addition to the recent randomized peer effects study Han and Li (2009). Once again, it is theoretically possible to exploit this non-linearity to improve total social welfare, but the exact implications of actually implementing such a policy are beyond the scope of this analysis.

2.5 Conclusion

This study describes persistence in the life and physical sciences at a large selective research university. Examining the role of preparation, grades and peers, I find a large impact of grades on persistence in both fields. While preparation is strongly correlated with persistence, there is little evidence that preparation directly impacts persistence outside of its impact on college grades.

As expected, students who receive higher grades in non-science courses are more likely to transfer out of the sciences and students who receive higher own field grades are less likely to transfer out of the sciences. This mechanism is very similar for life and physical science majors, except that life science majors who perform well in physical science courses are actually more likely to persist in the life sciences. This may reflect the fact that very few students transfer between life and physical science majors and the threat to major persistence is transferring to non-science majors.

The primary descriptive finding is that males and females persist equally well in the life sciences but a large gap exists in the physical sciences. This gap narrows considerably when controlling for other factors, but even controlled estimates show that females have slightly worse persistence rates in the physical sciences. The raw persistence gap for black students compared to white students is present in both the life and the physical sciences, but can largely be explained by performance and preparation factors. Conversely, even when controlling for other factors, Hispanic students are found to have much lower persistence rates in the life sciences than white students.

This study also documents evidence of peer effects in the physical sciences but finds no evidence of similar effects in the life sciences. For the physical sciences, exposure to peers who have a higher ex-ante probability of persistence is found to increase the probability of persistence. The impact of peers is shown to have important non-linearities where females and unlikely persisters experience the greatest gains from exposure to high quality peers.

Figures and Tables

Figure 2.1: Persistence rates for Life Sciences by Gender

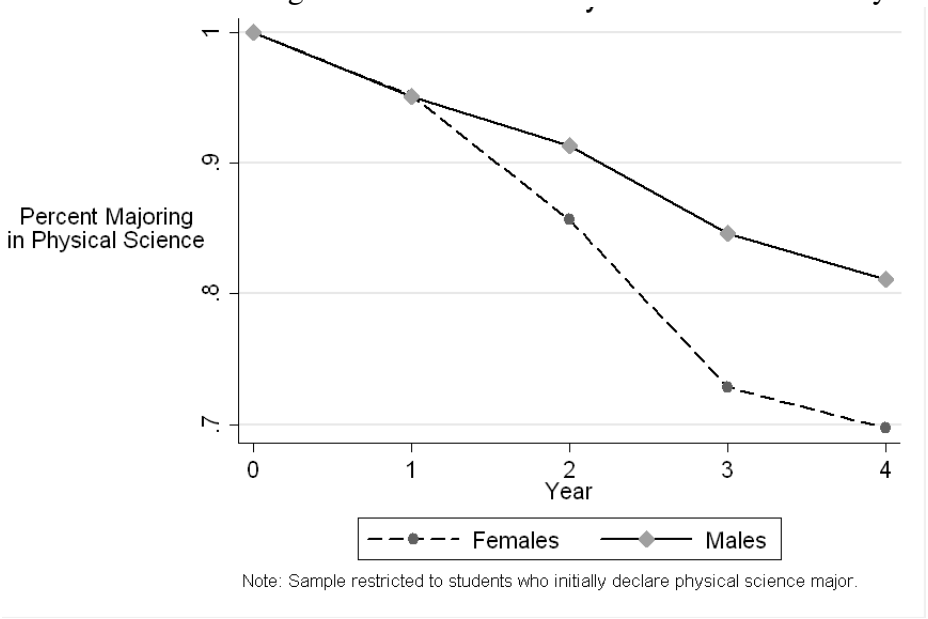


Figure 2.2: Persistence rates for Physical Sciences by Gender

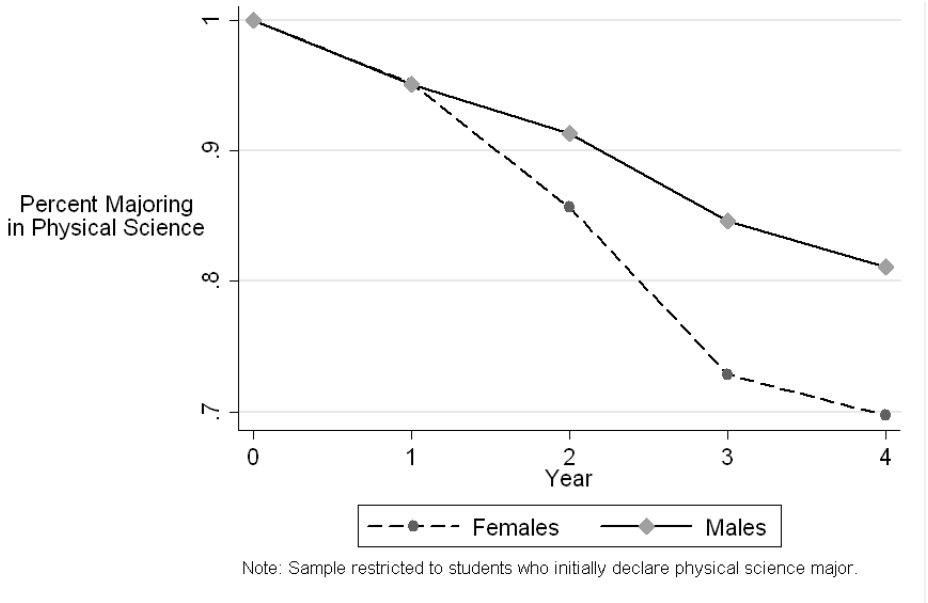


Figure 2.3: Persistence rates for Life Sciences by Race

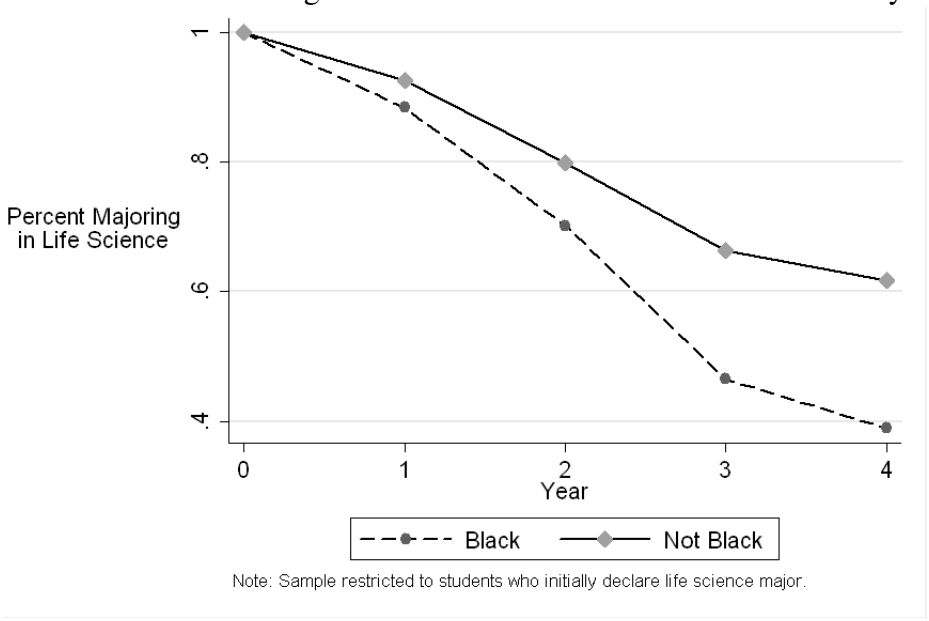


Figure 2.4: Persistence rates for Physical Sciences by Race

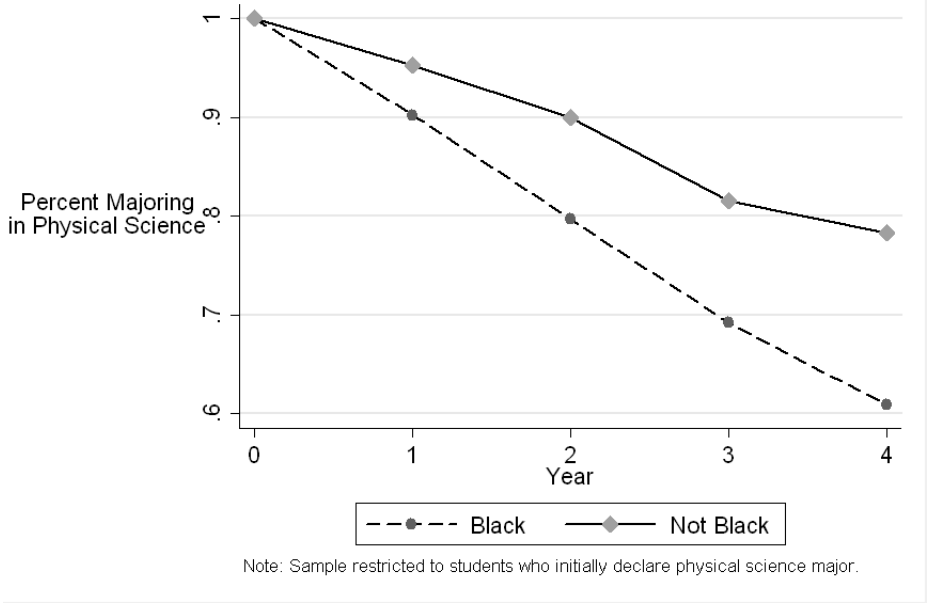


Figure 2.5: Persistence rates for Life Sciences by Race

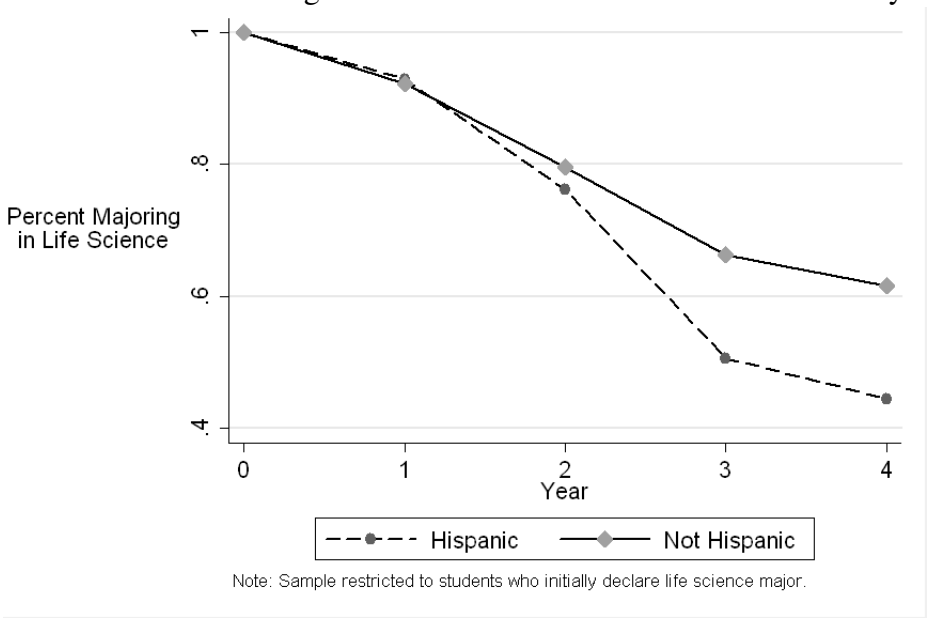


Figure 2.6: Persistence rates for Physical Sciences by Race

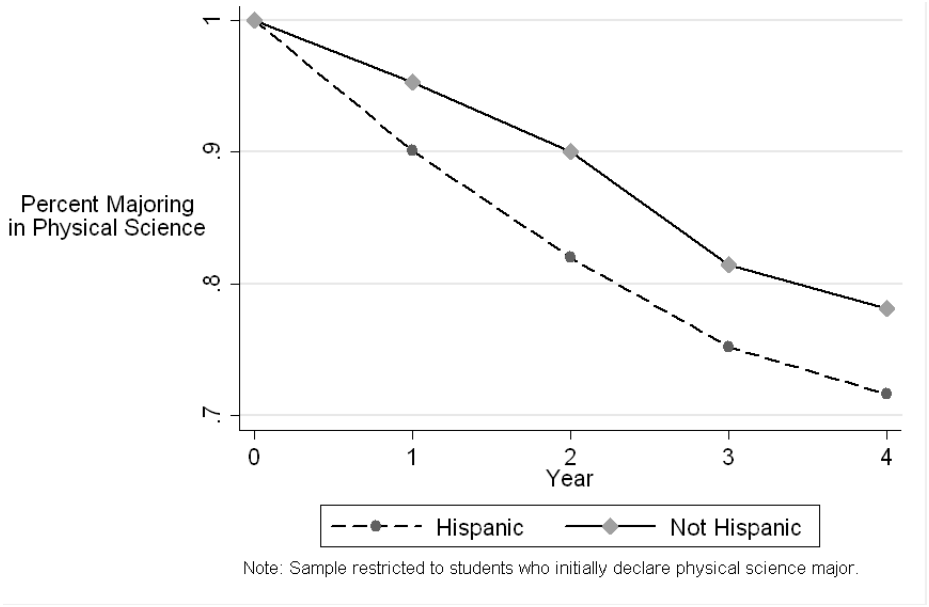


Figure 2.7: Persistence rates for Life Sciences by Race

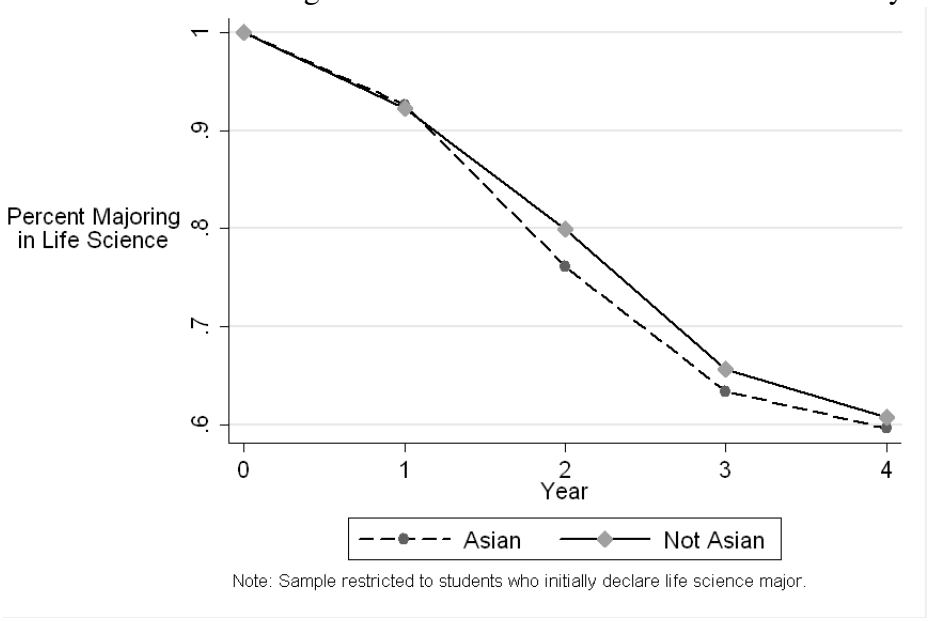


Figure 2.8: Persistence rates for Physical Sciences by Race

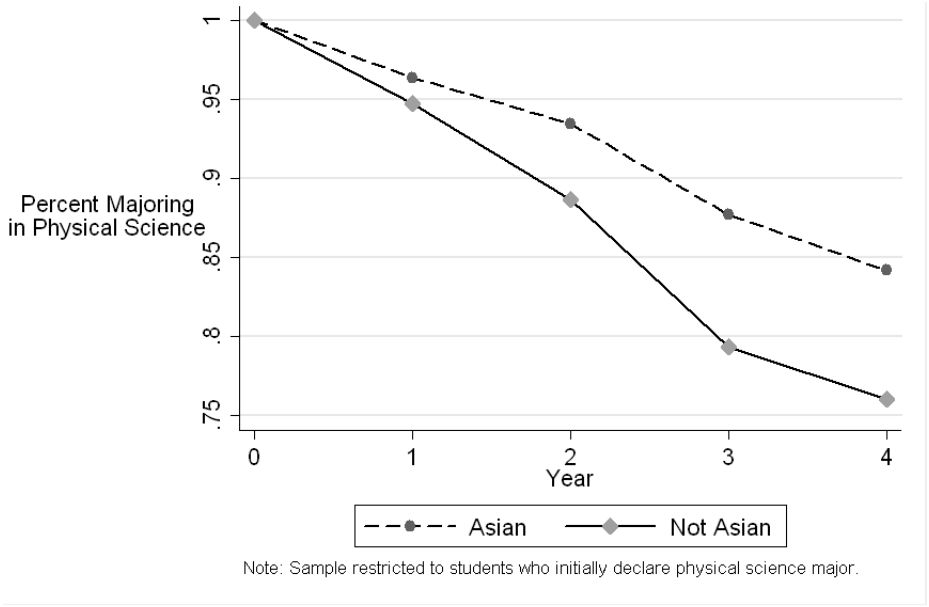


Figure 2.9: Persistence rates for Life Sciences by SAT

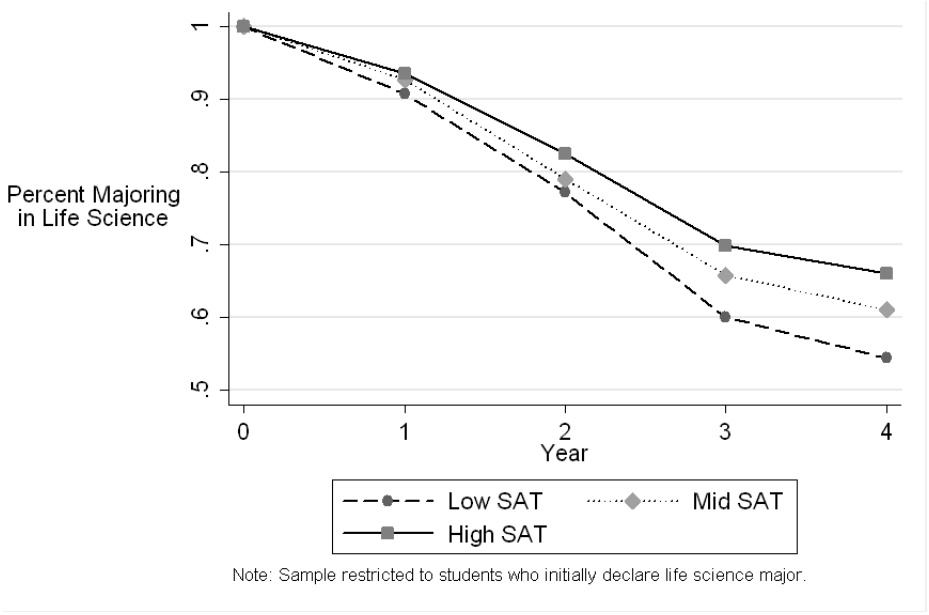


Figure 2.10: Persistence rates for Physical Sciences by SAT

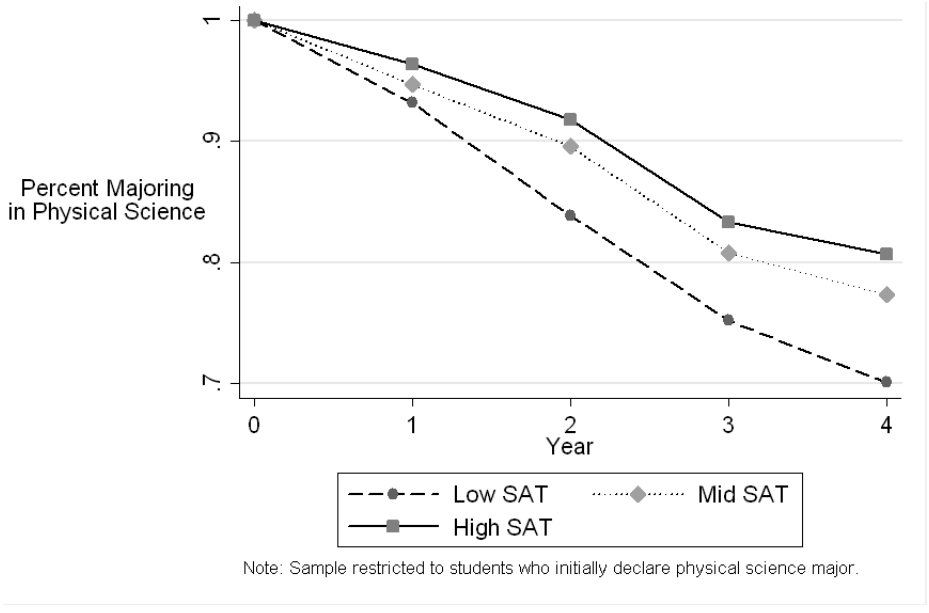


Table 2.1: Descriptive Statistics

		Intended Major			
	Mean	Not Science	Physical Science	Life Science	None declared
Female	0.487	0.570	0.287	0.613	0.597
Black	0.046	0.058	0.023	0.053	0.066
White	0.537	0.574	0.494	0.552	0.518
Asian	0.161	0.117	0.221	0.150	0.140
Hispanic	0.054	0.058	0.039	0.061	0.073
Multiple races reported	0.022	0.021	0.021	0.023	0.029
Native American	0.003	0.004	0.002	0.003	0.002
Race not reported	0.172	0.164	0.194	0.155	0.164
Number of incoming college courses	1.104	0.818	1.410	1.175	0.922
Incoming credit for calculus	0.237	0.153	0.341	0.229	0.196
Percentile (higher number is better rank)	0.938	0.922	0.949	0.949	0.926
SAT or equivalent	1358.242	1328.776	1398.941	1347.410	1349.001

Table 2.2: Intended Major

	Intended Major Field		
	Physical Science	Life Science	Non-Science
Female	-0.242***	0.134***	0.118***
	(0.011)	(0.007)	(0.011)
Black	0.009	0.020	-0.060**
	(0.024)	(0.017)	(0.022)
Hispanic	0.026	0.013	-0.074***
	(0.020)	(0.015)	(0.019)
Asian	0.094***	-0.024*	-0.108***
	(0.012)	(0.010)	(0.013)
Standardized SAT score	0.083***	-0.028***	-0.116***
	(0.006)	(0.004)	(0.006)
Standardized class percentile rank	-0.002	0.032***	-0.094***
	(0.007)	(0.005)	(0.008)
Incoming calculus credit	0.166***	-0.058***	-0.156***
	(0.015)	(0.012)	(0.015)
Number of incoming college courses	-0.027***	0.017***	0.014***
	(0.003)	(0.002)	(0.004)

* Significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors clustered at class level reported in parentheses.

Notes: Marginal effects from a multinomial logit are given in the table. The coefficients from all three columns come from a single regression on 15,508 observations. The variables "SAT score" and "class percentile rank" have been standardized to facilitate coefficient interpretation. For the variable "class percentile rank", lower numbers indicate a better ranking.

Table 2.3: Persistence in Physical Science

	(1) Coefficient (std. error)	Marginal	(2) Coefficient (std. error)	Marginal	(3) Coefficient (std. error)	Marginal
Female	-0.284***	-0.028	-0.275***	-0.027	-0.274***	-0.027
	(0.102)		(0.102)		(0.102)	
Black	-0.098	-0.010	-0.117	-0.011	-0.114	-0.011
	(0.274)		(0.274)		(0.277)	
Asian	0.416***	0.038	0.414***	0.037	0.404***	0.037
	(0.124)		(0.125)		(0.125)	
Hispanic	-0.172	-0.017	-0.165	-0.016	-0.168	-0.017
	(0.243)		(0.242)		(0.242)	
# of incoming college courses	-0.015	-0.001	-0.020	-0.002	-0.005	-0.000
	(0.037)		(0.037)		(0.039)	
Incoming calculus credit	0.172	0.016	0.199	0.019	0.298*	0.028
	(0.169)		(0.169)		(0.180)	
High school percentile	1.927	0.185	1.929	0.184	1.848	0.176
	(1.192)		(1.202)		(1.221)	
Standardized SAT score	-0.056	-0.005	-0.061	-0.006	-0.075	-0.007
	(0.066)		(0.067)		(0.068)	
GPA in 1st year life sci. courses	-0.219	-0.021	-0.220	-0.021	-0.218	-0.021
	(0.166)		(0.168)		(0.169)	
GPA in 1st year phys sci. courses	1.211***	0.117	1.201***	0.114	1.209***	0.115
	(0.149)		(0.149)		(0.153)	
GPA in 1st year courses	-0.359*	-0.035	-0.354*	-0.034	-0.368*	-0.035
	(0.203)		(0.202)		(0.205)	
Class size in core phys. sci. course	0.249***	0.024	0.074	0.007	0.102	0.010
	(0.035)		(0.075)		(0.073)	
Course fixed effect	No		Yes		Yes	
Cohort fixed effect	No		No		Yes	
Observations	19467		19467		19467	

* Significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors clustered at the student level reported in parentheses.

Notes: The dependent variable is major persistence to the fourth year in the physical sciences. The regression is restricted to students who intend to major in the physical sciences. Also included in the regression are missing indicators for cases where GPA or high school percentile is missing and peer characteristics.

Table 2.4: Persistence in Life Science

	(1) Coefficient (std. error)	Marginal	(2) Coefficient (std. error)	Marginal	(3) Coefficient (std. error)	Marginal
Female	0.021	0.004	0.011	0.002	0.001	0.000
	-0.102		(0.102)		(0.103)	
Black	-0.243	-0.046	-0.191	-0.035	-0.224	-0.041
	-0.218		(0.217)		(0.216)	
Asian	-0.346**	-0.065	-0.326**	-0.061	-0.342**	-0.063
	-0.147		(0.147)		(0.148)	
Hispanic	-0.425**	-0.082	-0.417**	-0.079	-0.441**	-0.083
	-0.204		(0.203)		(0.203)	
# of incoming college courses	-0.006	-0.001	-0.006	-0.001	0.018	0.003
	-0.036		(0.037)		(0.039)	
Incoming calculus credit	-0.255	-0.048	-0.225	-0.041	-0.170	-0.031
	-0.161		(0.163)		(0.165)	
High school percentile	-1.121	-0.204	-1.071	-0.193	-0.983	-0.175
	-0.983		(0.981)		(1.001)	
Standardized SAT score	-0.084	-0.015	-0.097*	-0.017	-0.108*	-0.019
	-0.058		(0.058)		(0.059)	
GPA in 1st year life sci. courses	0.621***	0.113	0.596***	0.107	0.599***	0.107
	-0.098		(0.098)		(0.098)	
GPA in 1st year phys. sci courses	0.324***	0.059	0.343***	0.062	0.359***	0.064
	-0.1		(0.100)		(0.101)	
GPA in 1st year courses	-0.442**	-0.080	-0.421**	-0.076	-0.456**	-0.081
	-0.178		(0.177)		(0.179)	
Class size in core life sci. course	-0.179***	-0.033	-0.160	-0.029	-0.151	-0.027
	-0.041		(0.165)		(0.168)	
Course fixed effect	No		Yes		Yes	
Cohort fixed effect	No		No		Yes	
Observations	6736		6736		6736	

* Significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors clustered at the student level reported in parentheses.

Notes: The dependent variable is major persistence to the fourth year in the life sciences. The regression is restricted to students who intend to major in the life sciences. Also included in the regression are missing indicators for cases where GPA or high school percentile is missing and peer characteristics.

Table 2.5: Impact of Average Peer Propensity Score

Panel A: Physical Sciences

	Course FE	Marginal	Course & Cohort FE	Marginal
All students	2.181***	0.208	2.160***	0.205
	(0.582)		(0.560)	
Bottom 25th percentile of propensity scores	2.889***	0.360	2.834***	0.353
	(0.986)		(0.954)	
Top 25th percentile of propensity scores	0.469	0.038	0.599	0.048
	(1.367)		(1.291)	
Females	2.968***	0.347	3.176***	0.370
	(0.975)		(0.927)	
Males	1.709**	0.146	1.600**	0.137
	(0.752)		(0.741)	

Panel B: Life Sciences

	Course FE	Marginal	Course & Cohort FE	Marginal
All students	-0.279	-0.050	0.396	0.071
	(0.767)		(0.782)	
Bottom 25th percentile of propensity scores	1.582	0.268	1.642	0.271
	(1.939)		(1.974)	
Top 25th percentile of propensity scores	0.264	0.043	2.273	0.361
	(1.514)		(1.593)	
Females	-0.201	-0.035	0.433	0.075
	(0.901)		(0.933)	
Males	-0.621	-0.114	0.272	0.050
	(1.525)		(1.559)	

Table 2.6: Grading Standards by Course Type and Student Intended Major

		Intended Major		
	All Students	Not Science	Physical Science	Life Science
Non-Science Course	3.33	3.31	3.35	3.39
Physical Science Course	3.13	3.08	3.16	3.06
Life Science Course	3.21	3.07	3.17	3.27

Table 2.7: Impact of Grades on Persistence for Males and Females

Panel A: Physical Sciences

	Females	Marginal	Males	Marginal
GPA in 1st year life sci. courses	0.034	0.004	-0.416*	-0.036
	(0.253)		(0.239)	
GPA in 1st year phys. sci. courses	1.150***	0.134	1.256***	0.107
	(0.250)		(0.197)	
GPA in 1st year courses	-0.608*	-0.071	-0.293	-0.025
	(0.352)		(0.257)	

Panel B: Life Sciences

	Females	Marginal	Males	Marginal
GPA in 1st year life sci. courses	0.617***	0.107	0.542***	0.099
	(0.124)		(0.167)	
GPA in 1st year phys. sci. courses	0.292**	0.050	0.521***	0.095
	(0.126)		(0.167)	
GPA in 1st year courses	-0.399*	-0.069	-0.547*	-0.100
	(0.225)		(0.286)	

Chapter 3: The Impact of Letter Grades on Student Course Selection and Major Choice: Evidence from a Regression-Discontinuity Design

3.1 Introduction

The extent to which students respond to their letter grades is crucial to understanding student major choice and course taking behavior. These decisions are of particular concern to policy makers given the considerable effort that has been devoted to improving major persistence, especially in the sciences. A common concern, first explicated in Sabot and Wakeman-Linn (1991), is that differential grading standards across the disciplines distort course taking behavior. In particular, students may decide to avoid a science major given the generally lower grades given in the sciences. Given that leniently graded fields are not necessarily more societally valuable than harsher grading fields, the distortion caused by student grade concerns has the potential to damage societal welfare.

Many studies have investigated whether students strongly respond to their letter grades. All such studies of which I am aware have found that students strongly respond to their letter grades such that students with higher letter grades in introductory courses are much more likely to major in that subject. The response of students to their letter grades is generally argued as efficient in that it promotes students sorting towards their comparative advantage; however, grading imbalances across fields distort this sorting process. Simulations from this literature suggest that equating letter grades across the university would have the (beneficial) impact of encouraging more students to pursue science. While research on this topic has spanned many years, institutions and disciplines, a fundamental obstacle to identifying the impact of letter grades on major choice is the possibility of unobserved factors which influence both major probabilities and introductory course letter grades. In particular, if students with more interest in a subject work harder, one would expect to see students with the highest performance also being the most likely to major. Though several studies have controlled for overall performance to identify a student's comparative advantage, this

approach does not address concerns that students put *relatively* more effort into the introductory courses in the field in which they plan to major.

Our study overcomes this obstacle by implementing a Regression Discontinuity (RD) design to identify the causal impact of letter grades on major choice and course performance. We supplement administrative records with a refined measure of course performance, collected directly from course instructors. This data allows us to observe not only the letter grade a student receives, but the exact numerical score they earned in the course. By comparing the major and course choices made by students with similar numerical scores, but different letter grades, I aim to identify the causal impact of the letter grades. To implement this analysis I collect original numerical scores for 65 introductory courses across 6 fields at a large selective research institution (LSRU). I combine these data with each student's full transcript, demographic information and major choices.

To examine this issue, I take two distinct approaches. In my first approach, I reproduce the typical analysis of the literature, as if I did not know exact numerical scores. I find evidence of a clear relationship between letter grades and major choices, which matches that found in the rest of the literature. When I add my collected numerical scores to this regression, however, I find that the entire correlation is explained by a linear function of numerical score. Once controlling for numerical score, none of the letter grades indicators are statistically significant and I am unable to reject the hypothesis that letter grades do not contribute to the model. In my second approach, I use the exact numerical score to implement an RD design testing whether students are more likely to major and take more course work in fields in which they earn higher grades. I find no evidence that students respond to their letter grades based on the RD specification and my estimates are fairly precise. Since cutoffs exist throughout the entire distribution, I am able to estimate a variety of local treatment effects and find no evidence that students respond to their letter grades whether at the top or bottom of the overall distribution. While I am unable to examine students who are not on a grade margin, the students who are of most interest to policy makers are precisely the students who are marginal and thus the RD research design is well suited to this application.

As in any RD design, my major concern is the possibility that students manipulate their scores

in order to fall just above a cutoff. This sort of manipulation is implausible on the final exam itself since students are unlikely to be able to precisely manipulate their test performance. However, it is very possible that extremely motivated students might effectively argue with their professor in an effort to boost their final letter grade, even when their exact numerical score does not qualify them for the higher grade. The key assumption made for the RD design is that the *numeric scores* are not manipulated to fall just above or below a cutoff—manipulation of letter grades will not bias estimates. To the extent that students are granted higher letter grades than their numerical score dictates, this simply converts the strict RD design to a fuzzy RD design. Importantly, even if the students who argue for higher grades are unobservably different than students who do not argue, the fuzzy RD design will yield consistent estimates. In order to ensure that the underlying course scores are not manipulated, I obtain the original spreadsheets used by professors in calculating numerical scores and confirm with the professors that these spreadsheets were not altered, even when a student successfully petitioned for a higher letter grade. I show that students frequently are granted higher letter grades than their numerical score dictates. However, I find no evidence of manipulation of the numerical scores themselves: the histograms of numeric scores around each letter grade cutoff show no evidence of scores humping just above grade cutoffs.

The plan of the text is as follows. I begin by reviewing the literature in Section 3.2 paying particular attention to the magnitudes found in previous research. Section 3.3 describes my data, Section 3.4 presents my first regression approach and Section 3.5 presents my RD approach. I provide a discussion of implications and how my work relates to previous research in Section 3.6 and conclude in Section 3.7.

3.2 Literature Review

A large literature has examined the determinants of major choice with particular emphasis on examining persistence in the sciences. Given the breadth of topics covered in this literature, I focus here on describing the literature that examines the role of grades in determining major and course

choices. Using data from Williams College, Sabot and Wakeman-Linn (1991) estimate how students respond to letter grades and examine how differential grade inflation across disciplines might distort major choice decisions. The authors find that controlling for performance in other subjects, receiving an A instead of a B in an introductory course increases the likelihood of taking a second course by approximately 10-20 percent for economics and English. Using a simulation, Sabot and Wakeman-Linn show that if economics graded as leniently as English at Williams College, enrollment in higher level economics courses would rise by 11.9 percent.

This basic point has been made repeatedly since that time and has been shown in a wide variety of disciplines and institutions. Christopher et al. (1994) examines the determinants of majoring and persisting in natural science and engineering at four highly selective institutions and similarly finds that letter grades are strongly correlated with declaring and remaining in these science majors. Similarly, Ost (2010) finds that students with a one point higher physical science GPA are 11 percentage points more likely to major in physical sciences and students with a one point higher life science GPA are 11 percentage points more likely to major in life science. Using data from a liberal arts college, Rask (2010) also finds that letter grades are important in predicting persistence in STEM fields such that a one letter grade change increases the probability of persisting by approximately five percentage points. Given that STEM departments grade more strictly than most departments in his study, Rask simulates the effect of equating grading standards across departments and concludes that this would increase STEM persistence by 2-4 percent.

In addition to discouraging persistence in STEM fields, student response to letter grades may explain racial or gender imbalances in certain majors. Rask and Tiefenthaler (2008) finds that economics students are sensitive to their grades in introductory courses and in particular, women appear more sensitive to these grades than men. Rask and Tiefenthaler posit that this sensitivity differential explains part of the gender imbalance in economics in higher level courses since women with equal performance to men leave economics at a higher rate. Owen (2010) confirms this finding for economics and finds that changing from a B to an A increases the probability of majoring by 15 to 20 percentage points among women while having no statistically significant impact for men.

While the literature examining the impact of introductory grades on course and major choice is well developed, the majority of the above studies rely on regression frameworks for identification. Several underlying behaviors are consistent with a strong correlation between letter grades and major choices and the regression framework is unable to distinguish between these underlying behaviors. First, it is possible that low letter grades in an introductory course cause students to leave a subject – either because they care about maintaining a high GPA or because they learn that their comparative advantage lies elsewhere. These two potential behavioral stories are intuitive and have been the primary interpretation of the literature. However, the relationship between major choice and introductory grades could also plausibly be generated by student response to underlying factors. In particular, students may choose to work hardest in the subject in which they intend to major, and as a result, they may earn their highest letter grades in their major fields. The policy implication of this phenomenon is very different. If students respond to their letter grades then equating average letter grades across departments has the potential to increase enrollments in initially low grading departments. If, on the other hand, students simply work hardest in their intended major, equating grading standards across departments will not have any direct impact on enrollment or major choice behavior.

The only study of which I am aware that is able to rule out an underlying factor and plausibly identify a causal impact of grades is Owen (2010). In her paper, Owen examines the impact of letter grades on major choice in economics using a similar RD methodology to the one used in my paper. She finds evidence of a strong impact of letter grades on major choices among women in economics and given her identification strategy these are interpreted causally. Given that Owen (2010) is the only paper that has estimated the causal impact of letter grades on major choices, I consider the replication of her analysis to be a contribution. This is particularly true because like many studies in this field, Owen (2010) focuses on a single institution and discipline and thus the results may not generalize to other settings.[1]

We extend Owen (2010) by considering a different institution and 6 disciplines. Also, in an

[1]Owen performs secondary analyses using a small liberal arts school, but the small sample at the second school prevents her from using a regression discontinuity design.

attempt to improve the precision of the estimates, I have collected more than ten times the number of observations as was used in the Owen. As a result, instead of using 30-60 observations on either side of the threshold, I am able to use nearly 1,000 students on either side of the threshold. The large amount of data facilitates breaking out the data more finely than previously possible and exploring interactions between grade responsiveness and factors such as financial aid status, gender, discipline and overall GPA. In the appendix, I attempt to replicate the exact analysis in Owen (2010). We are unable to replicate her findings, despite studying a similar institution and restricting my sample to just female economics students. Our large sample provides sufficient precision such that I am able to rule out the effect sizes found by Owen for my sample. We discuss potential reasons for this difference in results in the discussion section, but I am unable to provide a definitive explanation.

3.3 Data

The data used in this paper come from three distinct sources that are merged together. First, I collected grading spreadsheets from instructors at LSRU who teach large introductory courses. In collecting these data, when possible, I obtained the original spreadsheets that professors had used to record grades throughout the semester. In total I collected data from 65 course offerings across 6 disciplines. Due to confidentiality agreements made with specific instructors, I am unable to disclose the exact disciplines for certain subjects, and thus categorize courses as "Physical Science", "Life Science" or "Economics".[2] Two key pieces of information come from the grading spreadsheets. First, the spreadsheets include each student's final numerical score in a given course. Second, I carefully went through each spreadsheet and coded instances in which the professor indicated that he/she had altered a students numerical grade. The first key variable that records numerical scores is of central importance to my entire analysis while the second is useful

[2]Data was also collected for another social science discipline, but this is excluded from the main analyses because less than 1 percent of enrolled students intend to major in this subject. In practice, all results presented are robust to the inclusion of this subject, but estimates become less precise.

in assessing the extent to which grade manipulation might impact my results. Importantly, the data collected from instructors does not represent the universe of students at LSRU because it is restricted to only students who enrolled in one of the 65 course-offerings. In total, the spreadsheet data includes 20,774 students-course observations representing 9,565 students over a 11 year period (2000-2010).

Second, the registrar at LSRU provided the entire transcript for each student in the study population for the entire duration of their enrollment at LSRU. This data includes unique course identifiers and letter grades received for every course completed in addition to information on a students declared major(s).[3] From the transcript, I calculate cumulative GPA, semester GPA and categorize course taking behavior. Using a unique student identifier, this data is merged to admissions data from LSRU. The admissions data include basic demographic variables, financial aid information and SAT/ACT scores for each student. In addition, the admissions data include information on students' intended majors, which they list on their application for admission. The match rate between the three sources of data is very high for the years 2005-2010, but because LSRU changed administrative systems during the timeframe, I am unable to match all admissions variables prior to 2005. The 2000-2010 data has 20,334 student-course observations matched to transcripts and where possible, I use all of these observations. For some analyses, noted in the text, this sample is reduced as a result of missing admissions data in early years. The sample that focuses on 2005-2010 timeframe includes slightly over 13,000 observations.

The final merged dataset thus includes a complete course history for each student and two related measures of performance for the collected introductory courses. The first measure of performance is the exact numerical score the student received in the course (for example a 91/100). The second measure of performance is the letter grade from the student transcript, ranging from an F to an A+. These letter grades are converted to the LSRU GPA scale ranging from 0 to 4.3 where a B+ is a 3.3 rather than a $3.\bar{3}$ and an A- is a 3.7 rather than as $3.\bar{6}$. Throughout the remainder of

[3] If a student enrolls in a class but drops the course within the first several weeks, this course will not appear on the transcript or in my data. If a student drops the course after the designated drop period, I observe that student-course combination in my data.

the paper, I refer to these performance measures as numerical score and letter grade respectively.

Because different courses use different scales, the numeric scores are standardized to a 0 to 4.3 scale which is analogous to the 0 to 4.3 GPA scale but is measured continuously. This standardization makes across-course comparisons possible and also facilitates comparisons to the previous literature. In practice, this standardization is performed by mapping course grading cutoffs to the GPA scale and then mapping each student's score according to the distance from the cutoff. More exactly, I use the following formula, where γ_1 and γ_2 are the grade cutoffs in the original distribution, y is the student's percentage score in the course and α_1 and α_2 are the grade cutoffs being mapped to on the 0 to 4.3 scale.

$$Standardized\ Score = (y - \gamma_1)\frac{\alpha_2 - \alpha_1}{\gamma_2 - \gamma_1} + (\gamma_2 - \gamma_1) \tag{1}$$

For example, if a course initially grades on a 100 point scale where 97 or above is an A+ and 93 or above is an A, I map 97 to a 4.3 and map 93 to a 4.0. A student who received a 95 would be mapped to a 4.15 and a student who received a 96.4 would be mapped to a 4.255. While the GPA scale ranges from 0 to 4.3, the continuos version allows for some grades to exceed 4.3 since anyone who earns a numerical score above the A+ cutoff will be mapped to above a 4.3.

The first three columns of Table 3.2 show descriptive statistics for my data split by course discipline. Of the 2,072 students I observe taking introductory economics, 43 percent are female, 2 percent are black and 7 percent are hispanic. These demographic characteristics are fairly similar in engineering and the physical sciences but are dramatically different in the life sciences, where the gender imbalance is reversed and there is higher representation of black students. SAT scores (or ACT equivalents) are highest among students taking engineering and physical science courses and lowest among students taking life science courses; however, this pattern is not reflected in cumulative college GPA.

The most substantive difference between the three course categories is the intentions of students taking these courses. Nearly 70 percent of students taking engineering or physical science intro-

ductory courses intend to major in the course discipline. This is in stark contrast to the less than 5 percent of students taking economics who intend to major. The primary cause of this difference is the fact that students majoring in engineering are required to apply to the engineering school and list engineering as their intended major whereas there is no such requirement for economics majors (who enroll in the liberal arts portion of LSRU). Another potential reason for this difference is that introductory economics requires less technical background than do introductory engineering courses and thus students may be more likely to enroll in introductory economics purely out of topical interest. Of students who enroll in economics 17.3 percent choose to major in economics. The analogous figure is 60 percent for engineering and 53 percent for life sciences. This does not imply that engineers and life science courses have higher major persistence but simply reflect the fact that economics is a popular course among all students.

The last three columns of Table 3.2 restrict the attention to only students who eventually major in the course subject. These students are fairly similar to the other students in their classes with the notable exception that students who eventually major perform better in their introductory courses than students who do not major. Importantly, the demographic characteristics are similar between the students taking introductory courses and those majoring in the subject, suggesting that for this recent timeframe, persistence rates are similar for males and females. Compared to the average student taking an introductory course, a larger fraction of students who eventually major intended to major in that subject.

3.3.1 Data Issue: Imputing Grading Cutoffs

While my data is improved over previous research, one important limitation is that I do not exactly know the grading cutoffs used for the majority of the studied courses. Since knowing the grading cutoffs is crucial to my entire analysis, I put in considerable efforts to ensure that grading cutoffs are imputed accurately. Unlike many imputation procedures, it is not simply adequate to obtain an unbiased estimate of the cutoffs – I require that my imputation procedure perfectly and exactly obtains grading cutoffs. We are fairly confident that the imputation procedure that I use

meets this high standard. The imputation procedure involves a quantitative imputation followed by manually inspecting each course to ensure that the imputation is not driven by students with manipulated letter grades. The quantitative procedure chooses the cutoff for grade X according to the highest numerical grade received by an individual with a letter grade below X. In order to explain the complete imputation procedure it is useful to consider an example. Table 3.1 shows 10 scores from students around the B+/A- cutoff in a hypothetical course. Because each course used has hundreds of students, the density around any given cutoff is quite high and the example below is representative of the typical course in terms of density.

Table 3.1: Hypothetical Course

Student ID	Numeric Grade	Letter Grade
1	89.544	B+
2	89.662	B+
3	89.781	A-
4	89.824	B+
5	89.932	B+
6	90.031	A-
7	90.125	A-
8	90.132	A-
9	90.209	A-
10	90.311	A-

In the above example, the algorithm identifies student 5 as having the highest numerical grade of any student with below an A- letter grade. The imputed cutoff is then calculated as the average of that student with the next highest students score. In this case, averaging student 5, and student 6 yields and estimated cutoff of 89.9815. This imputation procedure is relatively simple, but performs exceptionally well. For the sample of courses for which I know the exact cutoffs, the imputation is typically within 0.02 points of the correct cutoff and always within 0.1 points of the correct cutoff. Once the grade cutoffs are imputed following the above procedure, I manually inspected each course to make sure that cutoffs appear appropriate and are not driven by students whose numeric grades were manipulated.

Note that in this example, student 3 received an A- but falls below imputed cutoff point. This

situation is common in my data and I attribute this phenomenon to either persistent students who argue for higher grades or generous professors who take into account motivation or performance trends in assigning letter grades. Importantly, I observe the original distribution of numerical scores, prior to the manipulation that results in student 3 receiving an A- and thus, this type of grade manipulation will not bias my estimates.

3.4 Regression Model

Before examining the evidence from the Regression Discontinuity model, I first consider how models used in the literature are altered when I include a control for numerical score. While a variety of models have been used to estimate the impact of grades on major choice, the key features of every model examines how course letter grades relate to major choice, conditional on general academic performance in other courses (Ost, 2010; Owen, 2010; Rask, 2010; Rask and Tiefenthaler, 2008; Sabot and Wakeman-Linn, 1991). We follow the literature in my approach and estimate the following model as a baseline.

$$Y_{it} = X_i\beta + Z_{it}\alpha + \sum_{j=D-}^{A+} \delta_j J_{it} + \gamma_j + \epsilon_{it} \tag{2}$$

Y_{it} is one of two measures of major choice. The first measure is an indicator for whether the student eventually majors in the relevant subject while the second measure is a count of the total number of credit hours taken in the relevant subject over the following three semesters.[4] X_i is a vector of time invariant characteristics including demographics, SAT score or ACT equivalent, and an indicator for whether the student listed the field as his/her intended major on the LSRU application. The vector Z_{it} includes cumulative GPA in time t, GPA in time t, and credit hours taken in time t. γ_j is a course fixed effect intended to capture important determinants of major

[4]Looking at course behavior over the following three semesters is motivated by a desire to smooth idiosyncratic course taking behavior driven by the availability of certain courses in only the spring or fall; however, all results presented in the paper are similar when looking at course taking behavior only in the semester immediately following the introductory course.

choice such as professor or peer quality (Carrell et al., 2010; Ost, 2010).

The key variables of interest are the coefficients on the dummy variables denoted by δ_j. Equation 2 is estimated as a linear probability model, but using a probit to predict major choice or a count model to predict subject credit hours yields similar results.

Equation 2 is the model typically estimated in the literature and the results for my sample are given in columns (1) of Table 3.3. Just as in the most papers in the literature, the results presented in the first and fourth columns of Table 3.3 paint a clear picture of the relationship between letter grades and major choice. Controlling for performance in other classes, students with better letter grades are more likely to major in the field and the magnitude of this difference is large. A student who receives an A- in an introductory course is 5 percentage points more likely to major in the subject than a student who receives a B+. Moving from an A- to a B- lowers the probability of majoring by nearly 9 percentage points and moving from an A- to a C- lowers the probability of majoring by over 17 percentage points. While these effect sizes are very large, they are very consistent with the rest of the literature that finds that, controlling for overall GPA, an increase of one point on a four point scale in one's introductory class is associated with a 15-20 percentage point change in the probability of majoring in the subject.

Column (4) of Table 3.3 shows the results from the same model when predicting the number of credit hours taken in the field in the following three semesters. This variable is intended to capture more nuanced variation in subject interest, but naturally, credit hours taken is correlated with eventual major choice. The results for credit hours are less consistent than for major choice and better letter grades are not monotonically associated with more credit hours. Lower letter grades in introductory courses are still generally associated with taking fewer subsequent credit hours and the impact is statistically significant when considering large letter grade changes. For example, students who receive a B- take 1.285 more credit hours than students who receive a C-.

While the relationship between letter grades and major choice is strong, whether this should be interpreted causally is unclear. It is possible that higher letter grades cause students to major in a subject, or it is plausible that students with the most interest or talent for a subject will both perform

well in their introductory course and subsequently choose to major. To distinguish between these two explanations, I add numerical score as an additional control that is intended to proxy for a student's natural talent or interest in a subject. Specifically, I estimate

$$Y_{it} = X_i\beta + Z_{it}\alpha + \sum_{j=D-}^{A+} \delta_j J_{it} + \gamma_j + \omega S_{it} + \epsilon_{it} \tag{3}$$

where S_{it} is a students numerical score for class t and all other variables are defined as in equation 2. If students actually respond to the letter grades that they receive, then one would expect the dummy variables to remain significant after the inclusion of the numerical score. Column (2) of Table 3.3 shows that the inclusion of the numerical score eliminates the correlation between letter grades and major choices. The relationship between letter grades and major choice is no longer monotonic, the coefficients are reduced by an order of magnitude and there are no statistically significant differences between a B+ and other letter grades.

An alternative test of the importance of letter grades is given by the incremental F-test comparing a model with numeric score *and* letter grade dummies to a model with just numeric score . Specifically, I first estimate

$$Y_{it} = X_i\beta + Z_{it}\alpha + \gamma_j + \omega S_{it} + \epsilon_{it} \tag{4}$$

where all variables are defined as in equation 3. We use the incremental F-test to examine whether the model given by equation 3 that includes the letter grade dummies contributes any explanatory power compared to the model given by equation 4 that excludes the letter grade dummies. This test is shown in the bottom panel of Table 3.3 and shows that adding letter grade dummies does not improve the model, when numeric score is already controlled for. Similarly, when using future credit hours as the outcome, the incremental F-test shows that adding letter grade dummies does not improve the model, when numeric score is already included.

In summary, I am able to replicate the findings of literature using a similar model, but these findings are not robust to the inclusion of the numerical score variable that I collected.

3.4.1 Analysis of Females

Several previous papers have noted that females may be more sensitive to grade feedback than males (Crocker and Major, 1989; Owen, 2010; Rask and Tiefenthaler, 2008; Seymour, 1995). In order to investigate this possibility I re-estimate equations 2 through 4 on only the females in my sample. Table 3.4 shows that results are fairly similar when focusing only on women. The relationship between letter grades and major persistence remains strong, though it is no longer entirely monotonic. Column 2 shows that once I control for numerical score, the relationship between letter grades and major choices is dramatically reduced in magnitude and is no longer statistically significant. As with the entire sample, female students with lower letter grades tend to take fewer credit hours in a subject that they perform poorly in; however, this relationship is not robust to the inclusion of the numeric score control. Once controlling for numeric score the dummy variable for earning an "A" is negative and marginally significant and the F-test rejects the hypothesis that the letter-grade dummy variables do not improve the model at the 10% level. However, the overall relationship is highly non-monotonic and does not show broad evidence in support of the notion that earning a higher letter grade increases the number of credit hours taken in the field. That being said, given that the initial relationship between future credit hours and grades is relatively weak among women, I find these results to be inconclusive regarding whether letter grades matter in determining course choice among women.

3.5 Evidence from Regression Discontinuity Design

Based on the regression analysis, I conclude that the relationship between letter grades and major choice is likely driven by an underlying continuous process. To test this further, I use a regression discontinuity (RD) design to test for a structural break around each grade cutoff.

3.5.1 Humping and Sorting

Given that RD estimates rely on comparability between students on either side of the threshold, a threat to identification occurs if students sort around the cutoff in a systematic and unobserved fashion. In the case of sorting around a grade cutoff, one might be especially concerned, because grade cutoffs are sometimes known ahead of time and students have a strong incentive to put in just enough effort for their numerical score to fall above a cutoff, or a student might argue with his or her professor to receive a higher grade even when the numerical score falls just below the cutoff (**?**). Furthermore, even if students are unable to successfully petition for higher grades, it is plausible that professors will artificially raise certain students' numerical score based on student interest and motivation, student improvement during the semester or extenuating circumstances. Whether driven by students or professors, this type of grade manipulation will likely generate a very specific humping pattern in the histogram or numerical scores – a pattern that can be tested for directly. If many students who should have received scores just below the cutoff receive scores just above the cutoff, this will result in a hump in the histogram just above the cutoff and a valley in the histogram just below the cutoff. If no such pattern is evident in the histogram then this provides compelling evidence that students are not systematically sorting around the cutoff.

Importantly, if a student receives a higher letter grade than their numerical score justifies, this by itself does not violate the RD assumption in any way. The assumption is not that every student with a score below the cutoff receives the lower grade, but rather that the scores themselves are not manipulated in order to fall just above or below the cutoff. At LSRU, professors maintain their own personal records in addition to reporting official grades to the university. As long as professors do not manipulate their own personal records, manipulation of the official grade will not invalidate the RD research design in this application. To determine the likelihood of professors manipulating their personal records, I spoke with each professor who provided us the data to directly discuss this issue. Our conversations suggest that the professors in my sample never change the numerical scores in their own records, but sometimes will change official letter grades based on student petitions or their own judgement. In any case, if professors do manipulate the raw numerical scores

(and then later lied to us about doing so), this has the potential to bias estimates and the direction of this bias is likely in favor of finding a larger impact of grades on major choices. Under the plausible assumption that those most likely to major in a subject are also most likely to have their numeric scores artificially raised, the RD estimates will confound inherent interest or motivation with letter grades and overstate the impact of grades.[5] If grading thresholds are set endogenously to the score distribution, this will not bias estimates so long as the threshold is set independently of a student's unobserved motivation or subject interest. For example, if a professor sets grading cutoffs by looking for "natural breaks" in the distribution, this is will generate a valley on either side of the threshold, but is unlikely to result in students being unobservably different on either side of that threshold. Regardless, if endogenous grading scales are used, this will be evident in the histograms, particularly if professors look for "natural breaks" to determine cutoffs.

Figure 3.1 shows the histogram of numerical scores centered around the B-/B cutoff, which is the modal score. Since sorting and manipulation might be masked by the standardization process, the only modification made in the histogram is subtracting the cutoff, which cannot alter the basic shape of the histogram. Figure 3.1 shows that this histogram of letter grades follow a bell shape, increasing up until B/B- and then decreasing. In order to look more precisely at humping, Figures 2(a) through 2(i) show a zoomed in version of Figure 3.1, with the histogram of numeric scores centered around each cutoff. Scores are reported on the original 0 to 100 scale, but are standardized so that the cutoff is at zero in each figure. The histogram is shown with a bin size of 0.2 percentage points, but the patterns are not sensitive to displaying other similarly small bin sizes. As shown in Figure 3.1 the histogram steadily increases for lower grades, peaks in the B range and then steadily decreases in the A range. Broadly, these histograms show no clear evidence of sorting around cutoffs, given that the histograms tend to move smoothly on either side of these cutoffs. The two histograms that are closest to exhibiting a humping pattern are Figure 2(a) around the D+/C- cutoff, Figure 2(g) around the B+/A- cutoff, and Figure 2(i) around the A/A+ cutoff. In

[5]We find it highly unlikely that the numerical grades in my sample have been manipulated both because the professors assured us that they were not and also because the professors have no incentive to manipulate their own records. The only grade that has any bearing is the official grade submitted to the university so I would expect that pressure to modify grades would be focused solely on this consequential variable.

these three figures, relative to the overall histogram trend, there appears to be slight humping to the left of the cutoff. This is somewhat surprising given that if humping were to occur, I would expect that students would be pushed just over the threshold, not artificially kept just under the threshold. Based on the mass of evidence from these histograms, combined with directly asking professors about manipulation, I conclude that there is no evidence of manipulation of the raw numerical scores.

3.5.2 First Stage

The RD design requires that the latent variable (numerical scores) impacts the treatment (letter grades) in a discontinuous fashion. To examine whether this assumption holds, I examine whether there is a discontinuous jump in the probability of receiving grade X around the numeric threshold for X. For example, Figure 3(a) plots the fraction of students receiving a letter grade above D- versus the student's standardized numerical score. It is clear from Figures 3(a) through 3(i) that there is a large discontinuous increase in the probability of receiving a grade as one's test score crosses the necessary threshold. These figures also show that while a large discontinuity exists, numerical scores do not perfectly dictate letter grades. As the numeric score approaches the cutoff, more students are bumped up to the higher grade such that just below the cutoff nearly 20% of students receive a higher letter grade than their numerical score dictates. Regardless, there remains a large discontinuity at the cutoff since nearly all students who receive a numerical score above the cutoff are given the higher letter grade. The fact that numerical scores do not perfectly dictate letter grades transforms my empirical approach from a strict RD to a fuzzy RD, but the intuition and implementation of the design is largely the same.

An alternative presentation of the same basic result is shown in Figure 3.4. This figure plots average letter grades (converted to a 0 to 4.3 scale) against average numeric score (standardized to the same scale). Each dot in this figure represents a bin of students who have a given numeric score. If numeric scores were perfectly predictive of letter grades, one would expect to see a perfect step graph where the letter grade jumps discontinuously at each cutoff and the average

letter grade in between each cutoff is constant. Figure 3.4 shows a pattern that is close to a stepwise pattern, but exhibits a very slight slope, particularly as numeric scores approach each cutoff. The discontinuities are very clear and are particularly stark for grades above a D+.

3.5.3 Second Stage

Given that letter grade assignment jumps discontinuously around grade cutoffs, if letter grades impact major choices, I expect that the fraction of students majoring in a subject will jump discontinuously around the grade cutoffs as well. As a first step, I simply plot the fraction of students majoring in the course subject against these students' numeric scores in the introductory course. Figure 3.5 shows the relationship between major choice and numeric scores. On this figure, the points that land on a vertical line correspond to students who just barely earned a numerical score at or above the grade cutoff. If the proportion of students majoring in a subject jumps discontinuously at each line, this would therefore be evidence that letter grades are impacting major choices. Instead, Figure 3.5 shows little evidence of discontinuous jumps at grading cutoffs. Only the 3.0 (B) cutoff shows a potential jump relative to trend, and the increased probability at 3.0 is not persistent as numeric scores rise above 3.0. Also, the discontinuity at 3.0 is of similar magnitude to other jumps that occur far away from grade cutoffs (for example near 2.5). On the whole, visual inspection of the relationship between numeric grades and major choice shows little evidence of discontinuous jumps which is striking when one compares this to Figure 3.4 which shows clear discontinuous jumps at every grade cutoff. The combination of Figures 3.4 and 3.5 paint a picture which is very consistent with the regression results previously presented – introductory performance is correlated with major choices, but the letter grades themselves do not appear to impact major choice.

The results are fairly similar when considering course choices in the three semesters following the introductory course. Figure 3.7 shows no evidence of a consistent jump in the number of subsequent credit hours taken as the numeric score crosses letter grade cutoffs.

To empirically estimate the magnitude of any potential discontinuities, I use local linear regres-

sion.

3.5.3.1 RD: Local Linear Regression

To estimate a local linear regression at each cutoff, I restrict the sample to within 0.25 points of each threshold and use a rectangular kernel; however, results shown are robust across a number of bandwidth choices and are not sensitive to the choice of kernel. Specifically I estimate:

$$Y_{it} = X_i\beta + Z_{it}\alpha + \gamma_j + \omega C_{it} + \delta A_{it} + \xi(C_{it})(A_{it}) + \epsilon_{it} \qquad for\ |C_{it}| < 0.25 \tag{5}$$

The variable, C_{it} is student i's standardized numeric score for course j with the relevant grade cutoff subtracted. The variable A_{it} is an indicator defined as $\mathbf{1}(C_{it} >= 0)$ and the interaction of C_{it} and A_{it} is included to allow the slope to vary on either side of the cutoff. The parameter of interest is δ, which is the estimated discontinuity. The linear model is fit to only points within 0.25 points of the cutoff, which ensures that no figure includes more than one potential discontinuity. Figures 7(a) through 7(i) show how major choices change around each cutoff. Each figure plots major choice conditional on covariates against numeric scores and also includes a note of the estimated discontinuity ($\hat{\delta}$) along with a standard error taken from estimating equation 5.[6] The lines on either side of the cutoff are graphed based on the coefficient estimates from equation 5 ($\hat{\omega}$ and $\hat{\xi}$), rather than from fitting a line to the conditional major choice variable.

Estimating equation 5 on the nine letter grade cutoffs yields no statistically significant estimates. Of the nine estimates, five are negative and four are positive, and none of the figures show visual evidence of a discontinuity. Furthermore, the point estimates are uniformly small and an order or magnitude less than earlier findings (Owen, 2010). While these results are generally robust across specification choices, some combinations of kernels and thresholds yield statistically significant discontinuities for certain thresholds; however, the statistically significant estimates are quite sensitive to specification and so I do not consider them to be strong evidence of a discontinuity. In results shown in the appendix, I similarly find no evidence of a discontinuity when focusing just

[6]The conditional major choice variable is the residual from a regression of major choice on covariates.

on females in economics as was done in Owen (2010).

Similar to my results for predicting major choice, I find little evidence that letter grades influence credit hours taken. Figures 7(a) through 7(i) show how conditional credit hours change around each cutoff. The discontinuity estimates noted on these figures are taken from estimating equation 5 using subject credit hours taken in the following three semesters as the dependent variable. As can be seen in these figures, four of the estimates are negative, five of the estimates are positive and none of the nine estimates are statistically significant. The estimated discontinuities shown should be interpreted as the causal impact of earning a score slightly below the threshold or the "intent to treat" (ITT). In order to obtain an estimate of the "Treatment on the Treated" (TOT) it is necessary to scale up these estimates by a factor of approximately 5/4. This accounts for the fact that the first stage discontinuity is only 0.8 since 20% of students just below the threshold receive the lower grade. Regardless of whether one considers the ITT or the TOT however, the effect magnitudes are small, statistically insignificant and inconsistent across cutoffs.

Given that there is no visual evidence of any discontinuities in Figures 3.5 or 3.7 and none of the local linear regressions yield statistically significant estimates, I conclude that the regression discontinuity design provides no evidence that major or course choices are influenced by letter grades.

3.6 Discussion

In their influential 1991 paper, Sabot and Wakeman-Linn develop a model of course choice in which students derive utility from learning, from their grades and from discounted future benefits. In their model, while students' human capital benefits from learning through their coursework, good grades themselves improve satisfaction. This notion of a direct benefit to higher grades has informed future research and is supported by theoretic intrinsic and extrinsic factors. In addition to contributing to a "warm glow of achievement", many extrinsic rewards such as graduate scholarships, academic honors and parental approval are direct functions of letter grades. The result

from this model implies that students will pursue subjects in which they are best suited to learn, but this optimal behavior can potentially be distorted by the direct incentive of letter grades if different fields have different grading functions. If student behavior is indeed distorted by letter grade considerations, then one might expect that two students with roughly the same level of learning, but different letter grades, would have a different probability of majoring in a field. We find little evidence that this is the case. Taken as a whole, I believe that the results from the regression discontinuity design combined with the regression analysis do not provide support for the notion that letter grades causally impact major or course choices.

This finding has a number of implications. First, it suggests that if students learn about their ability through their relative performance in their coursework, this learning is not informed by the ultimate letter grade earned in the course. Second, this weakens the confidence with which I can predict the implications of policies being considered at several institutions to equalize grades across disciplines. Simulations of the impact of letter grades assume a causal impact of letter grades on major choices and my findings provide some reason to be skeptical.

Owen (2010) finds very different results from my paper in that she finds a very large positive impact of grades on major choices for women in economics. In the appendix I show that even when I exactly follow her methodology and restrict my sample to hers, I find no evidence of a grading impact. There are several possible reasons that my results differ, but none are entirely satisfactory explanations. First, while Owen (2010) and my paper both examine highly selective research universities, these universities may have different institutional factors that impede or facilitate choosing an economics major. Given that neither Owen nor I am permitted to reveal the institution used, a direct comparison of these institutional factors is not possible. That being said, by comparing our descriptive statistics, it is clear that these two institutions are slightly different in terms of who takes introductory economics. In my sample approximately 17 percent of students in introductory economics proceed to major in the field whereas in Owen's sample, only 12 percent major in the field. The average grades in the two samples are comparable and in both Owen's sample and my own, 44 percent of the introductory economics course is female. Given that

both institutions are selective research universities in the Northeastern United States, it is possible that the impact of grades on major choices is institution or sample specific and therefore further replications at other universities are necessary to characterize the effect.

A second potential reason for the difference in results is that the institution Owen analyzes gives grades without plusses and minuses, thereby making sharper discontinuities. While this can potentially explain the difference in results for the regression discontinuity estimates, it is not a convincing explanation for why I find such different results in a simple regression setting. We find similarly large "effects" of letter grades when not controlling for numerical scores but in my sample, controlling for numeric score eliminates these effects whereas in Owen's sample, the effect of letter grades is robust to controlling for numerical score. Furthermore, the sum of my effects across all 9 grading thresholds is substantially smaller than the point estimate Owen finds for just the B/A threshold, suggesting that the difference in grading scales at the institution is unlikely to fully explain the difference in my results. Furthermore, Owen extends her analysis to a liberal arts college that uses a plus/minus grading system and she finds large effects, directly contradicting the notion that the grading scale alone explains our divergent findings.

In both Owen's and my study the true effect that grades have on average major choices is potentially understated because the samples are necessarily restricted to students who choose to enroll in an introductory course. While these students are the appropriate sample when considering the determinants of major attrition, they are not representative of students in general. In particular, one might expect that given that science and economics courses have a reputation of giving relatively low grades, only the students who are least responsive to course grades would elect to enroll in such a course. Although I find no evidence that these students respond to their letter grades by changing their course of study, it is very possible that certain students avoid enrolling in the first place due to a fear of low grades. Students at LSRU are likely well informed regarding average grades across disciplines since median grade reports are made public to the student body. If the knowledge of median grades results in only the least grade-sensitive students enrolling in low grading departments, this might explain why I find no effect of letter grades on major choices for

my sample. Bar et al. (2009) finds evidence that students responded to the introduction of public median letter grades at Cornell and to the extent that LSRU students in my time frame are similarly responsive, the entire impact of grades on major choices may occur through the initial decision of whether to enroll in the introductory courses.

The regression discontinuity design aims to obtain the causal impact of letter grades by comparing two students of similar ability and motivation who received different letter grades. An interesting alternative exploration is to isolate the unobserved portion by comparing students who earn identical numeric scores but earn different letter grades. As I argue, a student who earns a score just below a grade cutoff but receives the lower grade is likely unobservably different than a student who earns a score just below a grade cutoff and receives the higher grade. This latter group of students had their letter grades artificially raised and I view this as suggesting that the student either demonstrated promise or argued forcefully for the higher grade, either of which I expect to be correlated with a higher likelihood of major persistence. To test this theory, I add an indicator for whether the students grade was artificially raised to the model given by equation 4. Columns (1) and (3) of Table 3.5, however, show no evidence that students who are given a higher grade than their numerical score dictates are more likely to major or take more credit hours. Columns (2) and (4) similarly show that this relationship does not appear even when allowing for the effect to differ across the distribution of numerical scores. This result has two possible interpretations and I cannot distinguish between the two. First, it is possible that grade adjustments are made primarily for students with extenuating circumstances that are uncorrelated with interest in the major. Second, it is possible that students do not petition for higher grades differently in subjects in which they plan to major compared to other subjects. In other words, if certain students petition for higher grades in all their courses regardless of their majoring plans, this would result in no correlation between having one's grade raised and majoring in the field.

3.7 Conclusion

This paper examines the causal impact of letter grades on major and course choices. Contrary to the broader literature, I find no evidence that letter grades themselves raise the probability of persisting in a major. As in past research, I document a strong correlation between letter grades and major choices, but I find that this correlation is explained by a continuous underlying process, namely course performance. In other words, I find that students are more likely to major in a subject when they earn high scores in the introductory course, but students who just barely receive an "A" are no more likely to major than students who just barely miss the "A".

There exists a large grading gap across the disciplines such that students interested in science face a trade-off between taking coursework in their preferred field, and maximizing their GPA. If students strongly respond to these GPA incentives, this might discourage prospective scientists from pursuing that major. Previous research has found that students strongly respond to these GPA incentives and thus based on this literature, rigorous grading practices are likely losing the sciences prospective students. Using an RD design, I find no evidence that students respond to their letter grades, casting doubt as to whether policies aimed at equalizing grades across the disciplines will indeed have the effect predicted by the previous literature.

One important question is whether major attrition in the sciences is problematic at all. Science majors theoretically are beneficial to society because they produce positive externalities and improve our global competitiveness; however, it is probable that the students most likely to produce these positive externalities are also likely to have performed well in their courses. If poor grades cause students to leave the sciences, then relatively harsh grading standards might in some ways be beneficial as a screening device. For this reason, one might hope that students on the border of A/A+ do not respond to their letter grades, but students who perform at the bottom leave the major as a result of their low grades. Our results show no evidence of students responding to letter grades at either end of the performance distribution.

While I find no evidence of a causal impact of letter grades on major choice, it is important to note that my finding is opposed to the crux of the literature and in particular contradicts Owen

(2010), a paper that uses an RD methodology at a similar institution. Our results suggest that the causal impact of letter grades is not as universally supported as previously thought. Future research should extend the use of this RD methodology to other institutions in order to establish whether letter grades matter. To the extent that the effect varies across institutions, it would be useful to understand what institutional factors impact how students' major choices respond to letter grades.

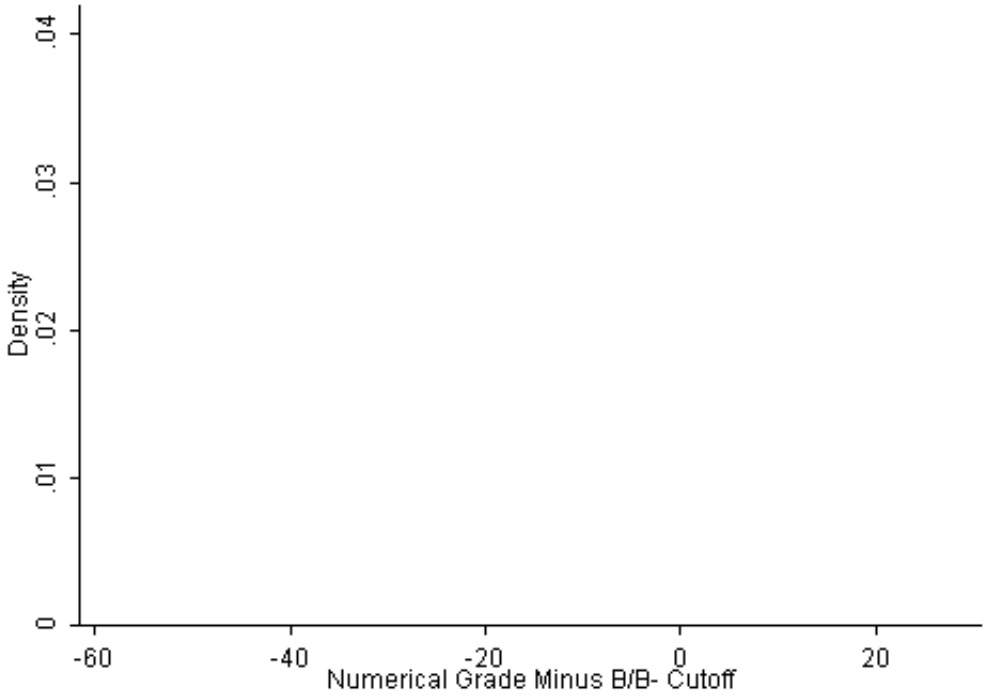
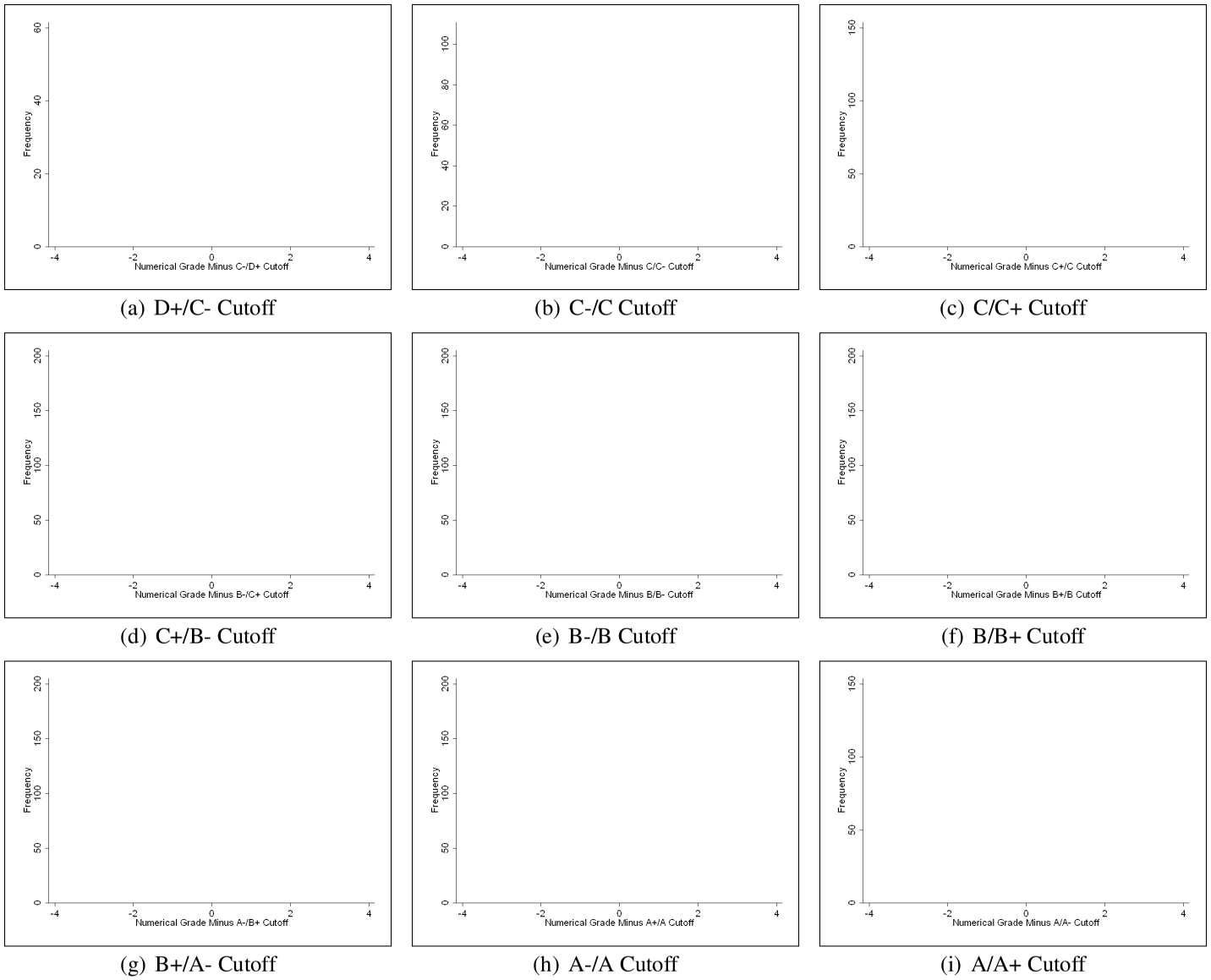


Figure 3.1: Histogram Normalized to B/B- Cutoff

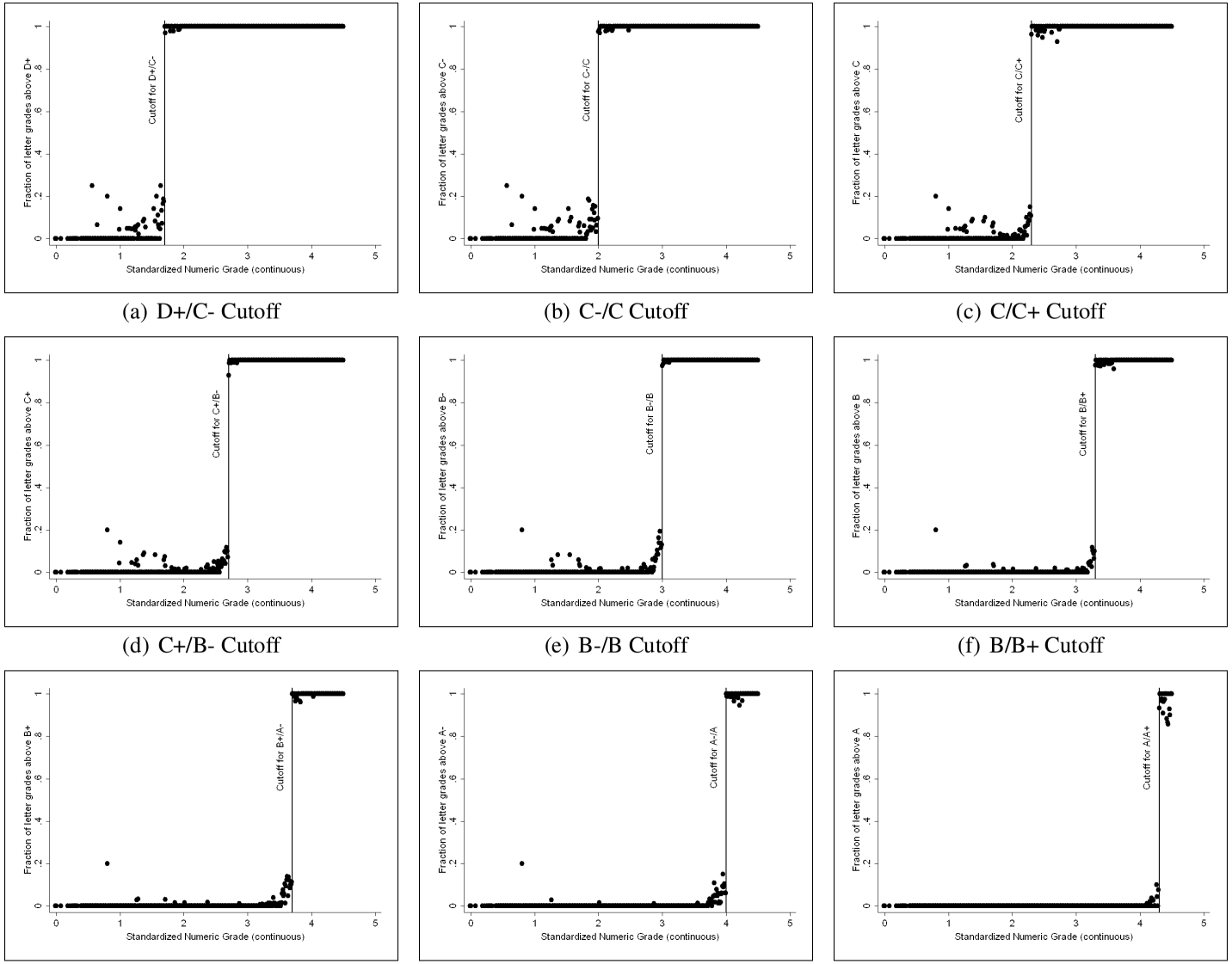


(a) D+/C- Cutoff (b) C-/C Cutoff (c) C/C+ Cutoff

(d) C+/B- Cutoff (e) B-/B Cutoff (f) B/B+ Cutoff

(g) B+/A- Cutoff (h) A-/A Cutoff (i) A/A+ Cutoff

Figure 3.2: Histograms Around Each Cutoff



(a) D+/C- Cutoff (b) C-/C Cutoff (c) C/C+ Cutoff

(d) C+/B- Cutoff (e) B-/B Cutoff (f) B/B+ Cutoff

(g) B+/A- Cutoff (h) A-/A Cutoff (i) A/A+ Cutoff

Figure 3.3: Discontinuity in Probability of Receiving a Given Grade Around Each Cutoff

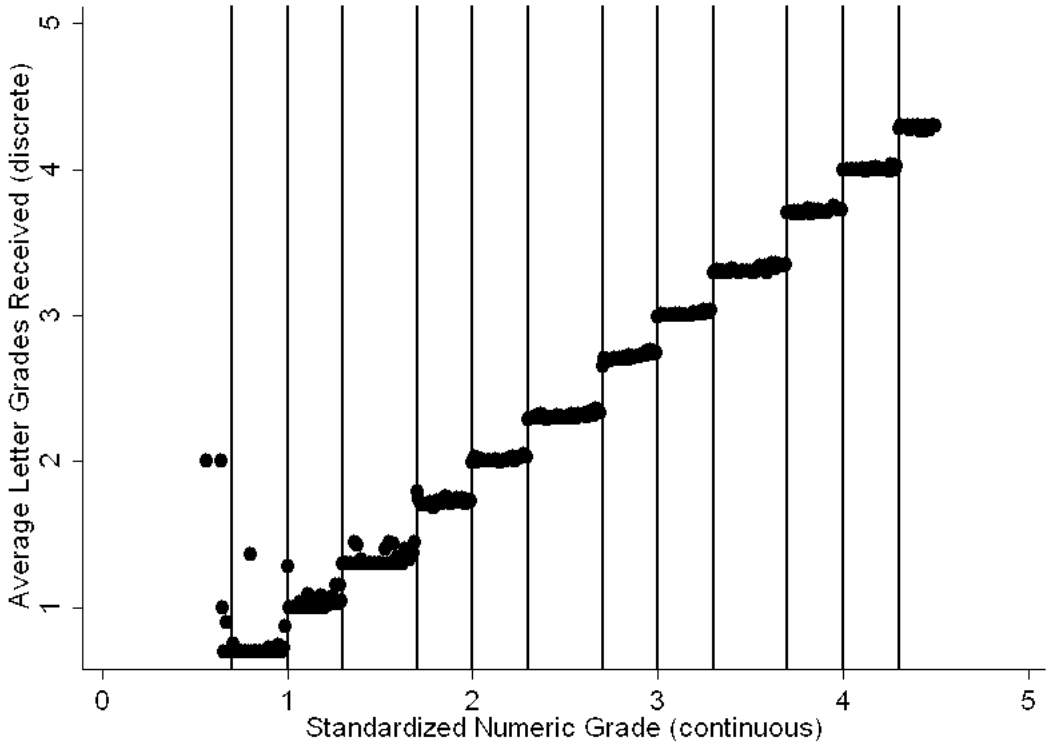


Figure 3.4: Average Letter Grades vs Numerical Score

Notes: Each point represents the average letter grade given to students in a given numerical score bin. The vertical lines show each cutoff value where 0.7, 1.0, 1.3, 1.7, 2.0, 2.3, 2.7 3.0, 3.3, 3.7 4.0 and 4.3 are the cutoffs for D-, D, D+, C-, C, C+, B-, B, B+,A-,A and A+ respectively.

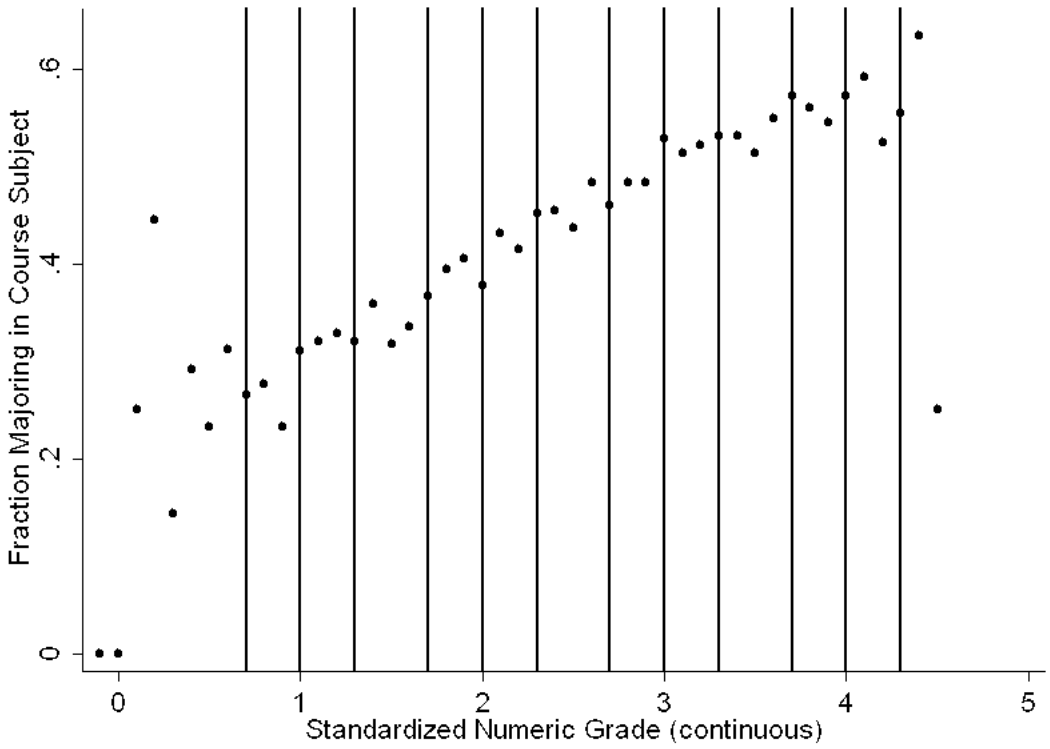


Figure 3.5: Fraction Majoring in Subject vs Numerical Score

Notes: Each point represents the fraction of students who major in the subject in a given numerical score bin. The vertical lines show each letter grade cutoff value where the cutoffs are the same as in Figure 3.4. Since the bin size is constant and the density is lowest at very high or very low scores, the variance is much larger at the extremes do to small sample sizes. The outliers at a numerical score of 4.5 and 0.4 represent very few students and thus these points should be interpreted with caution.

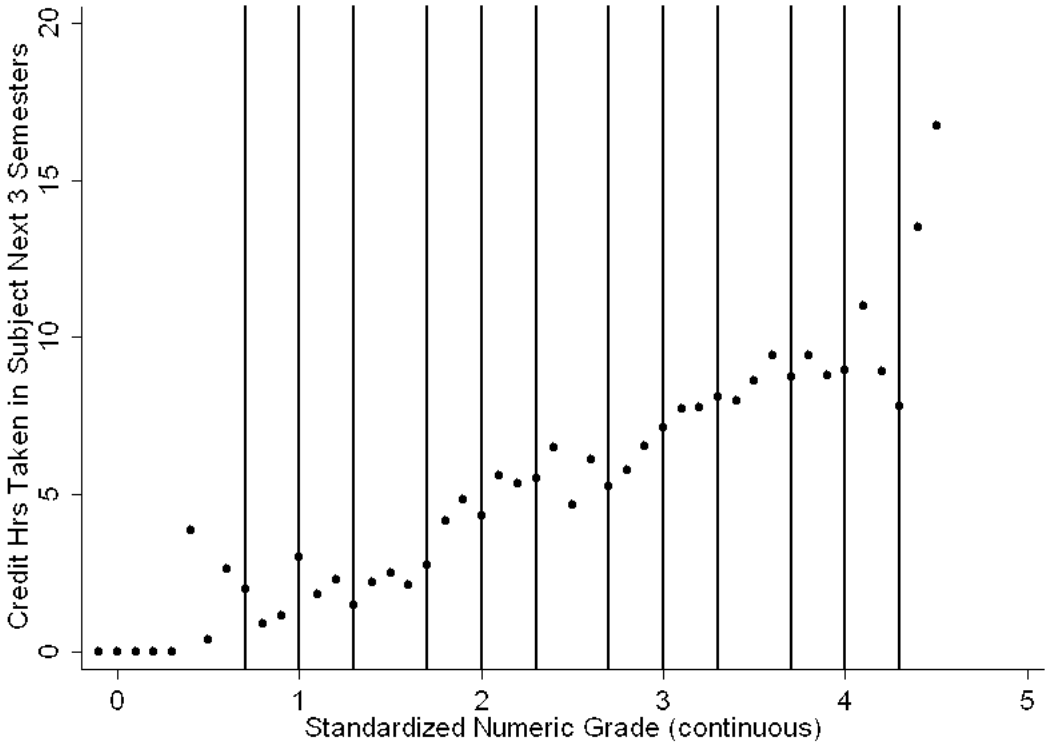
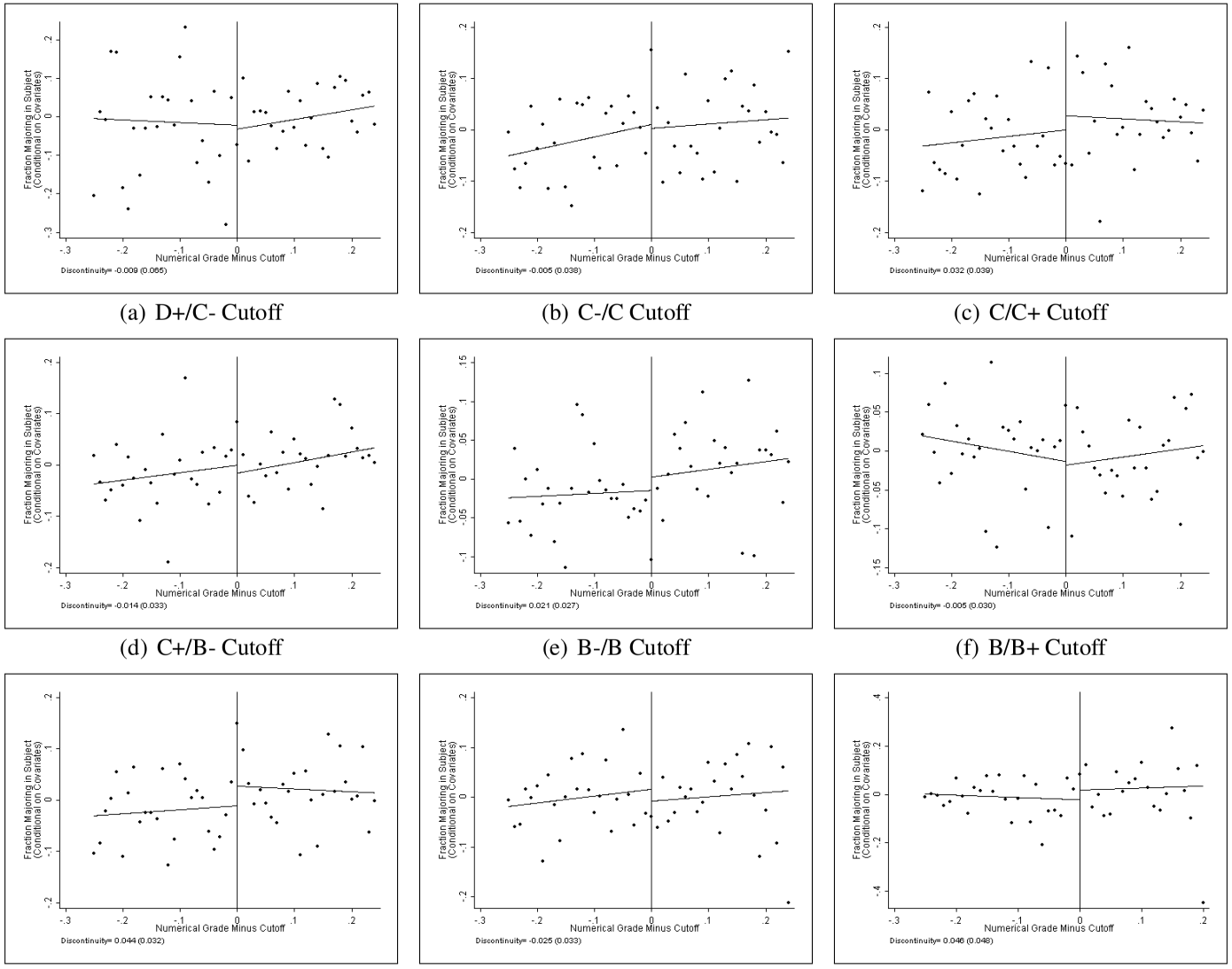


Figure 3.6: Credit Hours in Subject vs Numerical Score

Notes: Each point represents the fraction of students who major in the subject in a given numerical score bin. The vertical lines show each letter grade cutoff value where the cutoffs are the same as in Figure 3.4. Since the bin size is constant and the density is lowest at very high or very low scores, the variance is much larger at the extremes do to small sample sizes.

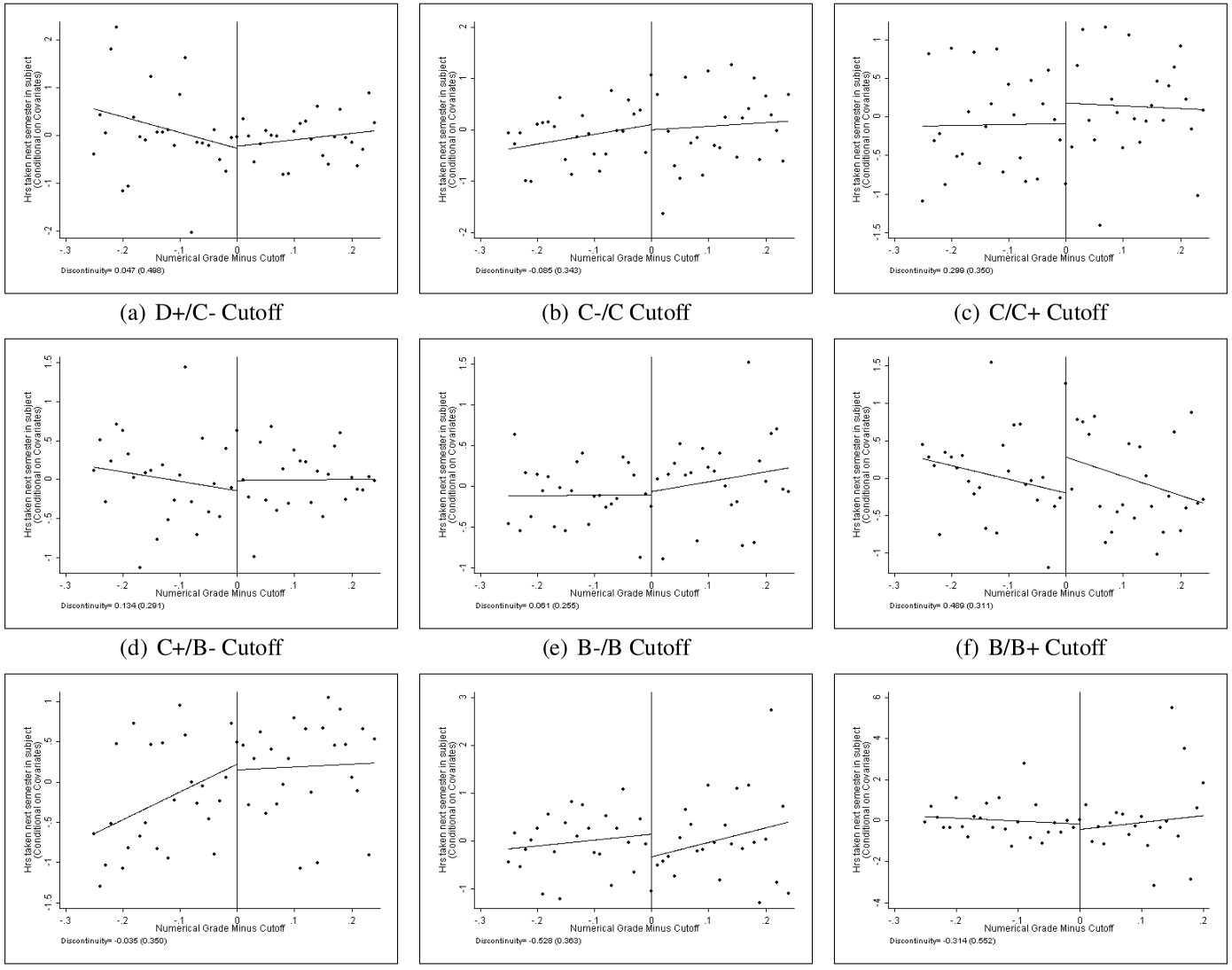


(a) D+/C- Cutoff (b) C-/C Cutoff (c) C/C+ Cutoff

(d) C+/B- Cutoff (e) B-/B Cutoff (f) B/B+ Cutoff

(g) B+/A- Cutoff (h) A-/A Cutoff (i) A/A+ Cutoff

Figure 3.7: Local Linear Regression Predicting Major Choice Around Each Cutoff



(a) D+/C- Cutoff (b) C-/C Cutoff (c) C/C+ Cutoff

(d) C+/B- Cutoff (e) B-/B Cutoff (f) B/B+ Cutoff

(g) B+/A- Cutoff (h) A-/A Cutoff (i) A/A+ Cutoff

Figure 3.8: Local Linear Regression Predicting Credit Hours Around Each Cutoff

Table 3.2: Descriptive Statistics

	Students Enrolled in Introductory Courses by Field			Restricted to Students who Eventually Major in the Introductory Course Subject		
Discipline	Economics	Engineering and Physical Science	Life Science	Economics	Engineering and Physical Science	Life Science
Major in Subject	0.17	0.59	0.53	1.00	1.00	1.00
Numeric Score (4.3 scale)	3.27	3.14	2.81	3.42	3.22	2.96
Intend to Major in Subject	0.03	0.69	0.36	0.16	0.84	0.61
Cumulative GPA	3.25	3.17	3.12	3.35	3.21	3.17
SAT or ACT equiv.	1393.12	1426.72	1386.17	1395.10	1429.27	1396.67
Female	0.43	0.41	0.60	0.44	0.40	0.57
Black	0.02	0.01	0.05	0.02	0.01	0.05
Hispanic	0.07	0.07	0.07	0.07	0.07	0.06
Observations	2,072	4,643	6,959	361	2,752	3,655

Table 3.3: Relationship between Letter Grades and Major and Course Choice

Dependent Variable:	Major in Subject			Credit Hrs in Subject During Following 3 Semesters		
	(1)	(2)	(3)	(4)	(5)	(6)
A plus	0.067*	0.008		-0.020	-1.028	
	(0.027)	(0.033)		(0.742)	(0.863)	
A	0.051**	0.009		-0.084	-0.802	
	(0.016)	(0.021)		(0.452)	(0.552)	
A minus	0.048**	0.025		0.724	0.337	
	(0.015)	(0.017)		(0.409)	(0.446)	
B	-0.011	0.013		-0.391	0.021	
	(0.013)	(0.015)		(0.342)	(0.392)	
B minus	-0.049***	-0.005		-0.442	0.308	
	(0.015)	(0.021)		(0.360)	(0.503)	
C plus	-0.081***	-0.014		-0.550	0.592	
	(0.016)	(0.027)		(0.391)	(0.653)	
C	-0.106***	-0.015		-1.045*	0.521	
	(0.018)	(0.035)		(0.443)	(0.842)	
C minus	-0.127***	-0.017		-1.285*	0.604	
	(0.020)	(0.042)		(0.501)	(1.006)	
Below C minus	-0.218***	-0.058		-2.215***	0.525	
	(0.023)	(0.058)		(0.532)	(1.340)	
Numerical score		0.072**	0.091***		1.226*	0.859***
		(0.024)	(0.008)		(0.567)	(0.191)
N	13,674	13,674	13,674	13,046	13,046	13,046
Incremental F-tests of whether dummy variables jointly contribute to model fit						
Incremental F-test:	Column (3) → (2)			Column (6)→ (5)		
	F statistic: 0.93 (p=0.498)			F statistic: 1.15 (p=0.324)		

* Significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors clustered at the student level reported in parentheses.

Note: The outcome is major choice in columns (1)-(3) and credit hours in column (4)-(6). All regressions also control for demographics, cumulative and current college GPA, credit hours taken contemporaneously with the courses analyzed, SAT score or ACT equivalent, an indicator for whether the student listed the major as their "intended major" on their application to LERU, and a course fixed effect. The omitted group for the letter grade dummies is B plus.

Table 3.4: Relationship between Letter Grades and Major and Course Choice (Only Females)

Dependent Variable:	Major in Subject			Credit Hrs in Subject During Following 3 Semesters		
	(1)	(2)	(3)	(4)	(5)	(6)
A plus	0.117**	0.058		1.176	0.371	
	(0.044)	(0.050)		(1.128)	(1.228)	
A	0.027	-0.015		-0.882	-1.450*	
	(0.024)	(0.030)		(0.569)	(0.649)	
A minus	0.063**	0.040		-0.102	-0.413	
	(0.022)	(0.024)		(0.534)	(0.561)	
B	-0.019	0.006		-0.745	-0.417	
	(0.019)	(0.021)		(0.426)	(0.476)	
B minus	-0.071***	-0.027		-1.332**	-0.732	
	(0.020)	(0.027)		(0.458)	(0.607)	
C plus	-0.079***	-0.012		-1.093*	-0.179	
	(0.022)	(0.035)		(0.475)	(0.739)	
C	-0.106***	-0.013		-1.231*	0.022	
	(0.024)	(0.044)		(0.533)	(0.925)	
C minus	-0.136***	-0.025		-2.571***	-1.060	
	(0.028)	(0.053)		(0.629)	(1.131)	
Below C minus	-0.216***	-0.054		-2.732***	-0.547	
	(0.031)	(0.072)		(0.626)	(1.463)	
Numerical score		0.073*	0.093***		0.984	0.936***
		(0.030)	(0.011)		(0.616)	(0.228)
N	6,953	6,953	6,953	6,696	6,696	6,696

Incremental F-tests of whether dummy variables jointly contribute to model fit

Incremental F-test:	Column (3) → (2)	Column (6)→ (5)
	F statistic: 1.27 (p=0.249)	F statistic: 1.75 (p=0.072)

* Significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors clustered at the student level reported in parentheses.

Note: The outcome is major choice in columns (1)-(3) and credit hours in column (4)-(6). All regressions also control for demographics, cumulative and current college GPA, credit hours taken contemporaneously with the courses analyzed, SAT score or ACT equivalent, an indicator for whether the student listed the major as their "intended major" on their application to LERU, and a course fixed effect. The omitted group for the letter grade dummies is B plus. The entire table is restricted to women.

Table 3.5: Relationship Between Unobservables and Major and Course Choice

Dependent Variable:	Major in Subject		Credit Hrs in Subject During Following 3 Semesters	
	(1)	(2)	(3)	(4)
Numerical Score	0.091***	0.092***	0.856***	0.869***
	(0.008)	(0.008)	(0.192)	(0.193)
Letter Grade Was Raised	0.002	0.027	-0.130	0.966
	(0.022)	(0.077)	(0.404)	(1.617)
Letter Grade Was Raised x Numerical Score		-0.009		-0.395
		(0.027)		(0.589)
N	13,674	13,674	13,046	13,046

* Significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors clustered at the student level reported in parentheses.

REFERENCES

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. Teachers and student achievement in the chicago public high schools. *Journal of Labor Economics*, *25*, 95–135.

Anderson, T. W., and Cheng Hsiao. 1981. Estimation of dynamic models with error components. *Journal of the American Statistical Association*.

Aronson, J, and M Inzlicht. 2004. The ups and downs of attributional ambiguity. *Psychological science*, *15*(12), 829–836.

Bar, Talia, Vrinda Kadiyali, and Asaf Zussman. 2009. Grade information and grade inflation: The cornell experiment. *Journal of Economic Perspectives*, *23*(3), 93–108.

Becker, Gary. 1964. *Human capital: a theoretical and empirical analysis with special reference to education*. University of Chicago Press.

Boyd, Donald, Pam Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2008. Who leaves? teacher attrition and student achievement. NBER Working Papers 14022, National Bureau of Economic Research, Inc.

Carrell, SE, RL Fullerton, and JE West. 2009. Does your cohort matter? measuring peer effects in college achievement. *Journal of Labor Economics*, *27*(3), 439–464.

Carrell, S.E., M.E. Page, and J.E. West. 2010. Sex and science: How professor gender perpetuates the gender gap*. *Quarterly Journal of Economics*, *125*(3), 1101–1144.

Carrell, SE, NBI Sacerdote, and JE West. Unpublished. Beware of economists bearing reduced forms? an experiment in how not to improve student outcomes.

Carrington, William J.. 1993. Wage losses for displaced workers: is it really the firm that matters? *Journal of Human Resources*, *28*(3), 435–462.

Christopher, Strenta A., R. Elliott, R. Adair, M. Matier, and J. Scott. 1994. Choosing and leaving science in highly selective institutions. *Research in Higher Education*, *35*(5), 513–547.

Clement, Michael B., Lisa Koonce, and Thomas J. Lopez. 2007. The roles of task-specific forecasting experience and innate ability in understanding analyst forecasting performance. *Journal of Accounting and Economics*, *44*(3), 378–398.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2006. Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources*, *41*(4), 778.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2007. Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, *26*(6), 673–682.

Crocker, J, and B Major. 1989. Social stigma and self-esteem: The self-protective properties of stigma. *Psychological review*, *96*(4), 608–630.

Dee, Thomas S.. 2005. A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review*, *95*(2), 158–165.

Eagly, A. 1978. Sex differences in influenceability. *Psychological Bulletin*, *85*, 86–116.

Foster, Gigi. 2006. It's not your peers, and it's not your friends: Some progress toward understanding the educational peer effect mechanism. *Journal of Public Economics*, *90*(8-9), 1455–1475.

Fournier, Gary M., and Tim R. Sass. 2000. Take my course, please: The effects of the principles experience on student curriculum choice. *Journal of Economic Education*, *31*(4), 323–339.

Gathmann, Christina, and Uta Schönberg. 2010. How general is human capital? a taskbased approach. *Journal of Labor Economics*, *28*(1).

Gibbons, Robert, and Michael Waldman. 2004. Task-specific human capital. *American Economic Review*, *94*(2), 203–207.

Goldhaber, Dan, Betheny Gross, and Daniel Player. 2007. Are public schools really losing their "best"? CRPE working paper 20072, Center on Reinventing Public Education.

Han, L, and T Li. 2009. The gender difference of peer influence in higher education. *Economics of Education Review*, *28*(1), 129–134.

Hanushek, Eric A., John F. Kain, Daniel M. O'Brien, and Steven G. Rivkin. 2005. The market for teacher quality. NBER Working Papers 11154, National Bureau of Economic Research, Inc.

Harris, Douglas N., and Tim R Sass. 2007. Teacher training, teacher quality and student achievement. CALDER Working Papers 3, CALDER.

Hoxby, Caroline, and Sonali Murarka. 2009. Charter schools in new york city: Who enrolls and how they affect their students' achievement. NBER Working Papers 14852, National Bureau of Economic Research, Inc.

Hoxby, Caroline M.. 2000. The effects of class size on student achievement: New evidence from population variation. *Quarterly Journal of Economics*, *115*(4), 1239–1285.

Jackson, Clement Kirabo, and Elias Bruegmann. 2009. Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics*, *1*(4), 85–108.

Jovanovic, Boyan. 1979. Job matching and the theory of turnover. *The Journal of Political Economy*, *87*(5(1)), 972–990.

Kambourov, Gueorgui, and Iourii Manovskii. 2009. Occupational specificity of human capital. *International Economic Review*, *50*(1), 63–115.

Kane, Thomas J., and Douglas O. Staiger. 2008. Are teacher-level value-added estimates biased? an experimental validation of non-experimental estimates. Nber working papers, National Bureau of Economic Research, Inc.

Kane, Thomas J., Douglas O. Staiger, and Stephanie K. Riegg. 2006. School quality, neighborhoods, and housing prices. *American Law and Economics Review*, *8*(2), 183–212.

Kletzer, Lori Gladstein. 1989. Returns to seniority after permanent job loss. *The American Economic Review*, *79*(3), 536–543.

Koedel, Cory, and Julian Betts. 2007. Re-examining the role of teacher quality in the educational production function. Working Papers 0708, Department of Economics, University of Missouri.

Koedel, Cory, and Julian Betts. 2008. Value-added to what? how a ceiling in the testing instrument influences value-added estimation. Working Papers 0807, Department of Economics, University of Missouri.

Lazear, Edward P. 1979. Why is there mandatory retirement? *Journal of Political Economy*, *87*(6), 1261–84.

Manski, Charles. 1993. Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, *60*(3), 531–542.

Neal, Derek. 1995. Industry-specific human capital: Evidence from displaced workers. *Journal of labor Economics*, *13*(4), 653–677.

North Carolina Department of Education. 2009. Accountability services. http://www.dpi.state.nc.us/accountability/testing/eog/.

Ost, B. 2010. The role of peers and grades in determining major persistence in the sciences. *Economics of Education Review*.

Owen, Ann. 2010. Grades, gender, and encouragement: A regression discontinuity analysis. *The Journal of Economic Education*, *41*(3), 217–234.

Owen, Ann L.. Forthcoming. Grades, Gender, and Encouragement: A Regression Discontinuity Analysis. *Journal of Economic Education*.

Parent, Daniel. 2000. Industry-specific capital and the wage profile: Evidence from the national longitudinal survey of youth and the panel study of income dynamics. *Journal of Labor Economics*, *18*(2), 306–323.

Poletaev, Maxim, and Chris Robinson. 2008. Human capital specificity: Evidence from the dictionary of occupational titles and displaced worker surveys, 1984–2000. *Journal of Labor Economics*, *26*(3).

Rask, Kevin. 2010. Attrition in stem fields at a liberal arts college: The importance of grades and pre-collegiate preferences. *Economics of Education Review*.

Rask, Kevin, and Jill Tiefenthaler. 2008. The role of grade sensitivity in explaining the gender imbalance in undergraduate economics. *Economics of Education Review*, *27*(6), 676–687.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458.

Rockoff, Jonah E.. 2004. The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*(2), 247–252.

Rothstein, Jesse. 2009. Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, *4*(4), 537–571.

Rothstein, Jesse. 2010. Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, *125*(1), 175–214.

Sabot, Richard, and John Wakeman-Linn. 1991. Grade inflation and course choice. *Journal of Economic Perspectives*, *5*(1), 159–70.

Sacerdote, Bruce. 2001. Peer effects with random assignment: Results for dartmouth roommates. *The Quarterly Journal of Economics*, *116*(2), 681–704.

Schanzenbach, Diane. 2007. What have researchers learned from project star? *Brookings Papers on Education Policy*.

Seymour, E.. 1995. The loss of women from science, mathematics, and engineering undergraduate majors: An explanatory account. *Science Education*, *79*(4), 437–473.

Stinebrickner, Ralph, and Todd R. Stinebrickner. 2006. What can be learned about peer effects using college roommates? evidence from new survey data and students from disadvantaged backgrounds. *Journal of Public Economics*, *90*(8-9), 1435–1454.

Stinebrickner, Todd R., and Ralph Stinebrickner. 2009. Learning about academic ability and the college drop-out decision. NBER Working Papers 14810, National Bureau of Economic Research, Inc.

Todd, Petra E., and Kenneth I. Wolpin. 2003. On the specification and estimation of the production function for cognitive achievement. *Economic Journal*, *113*(485), F3–F33.

Topel, Robert. 1991. Specific capital, mobility, and wages: Wages rise with job seniority. *Journal of Political Economy*, *99*(1), 145–176.

Turner, Sarah E., and William G. Bowen. 1999. Choice of major: The changing (unchanging) gender gap. *Industrial and Labor Relations Review*, *52*(2), 289–313.

Zimmerman, David J.. 2003. Peer effects in academic outcomes: Evidence from a natural experiment. *The Review of Economics and Statistics*, *85*(1), 9–23.